

Digitized by the Internet Archive
in 2020 with funding from
Kahle/Austin Foundation

Journal of Applied Psychology

KENNETH E. CLARK, Editor
University of Rochester

Consulting Editors

GEORGE E. BRIGGS, *Ohio State University*
NORMAN FREDERIKSEN, *Educational Testing Service*

LEONARD D. GOODSTEIN, *University of Iowa*
EDWIN R. HENRY, *Peace Corps*

JOHN HOLLAND, *American College Testing Program*

CLIFFORD E. JURGENSEN, *Minneapolis Gas Company*

LAURENCE S. MCGAUGHRAN, *University of Houston*

QUINN MCNEMAR, *Stanford University*

HAROLD F. ROTHE, *Beloit Corporation*

THOMAS A. RYAN, *Cornell University*

ALEXANDER G. WESMAN, *Psychological Corporation*

CLARK L. WILSON, *Batten, Barton, Durstine, and Osborn, Incorporated*

Volume 47, 1963

HELEN ORR
Managing Editor

ELIZABETH S. REED
Advertising Manager

VIRGINIA RICHARDS
Editorial Assistant

Published bimonthly by the American Psychological Association, Inc.
Prince and Lemon Sts., Lancaster, Pa. 17604 and 1333 16th St. N.W.
Washington, D. C. 20036

Second-class postage paid at Lancaster, Pa.

© 1963 by the American Psychological Association, Inc.

Contents of Volume 47

Andersen, L. Bryce, and Spencer, Patricia A. Personal Adjustment and Academic Predictability among College Freshmen	97
Arrowood, A. John. See Latané, Bibb.	
Baker, Robert A. See Osborn, William C.	
Bass, Alan R. See Triandis, Harry C.	
Bell, Forest O., Hoff, Alvin L., and Hoyt, Kenneth B. A Comparison of Three Approaches to Criterion Measurement	416
Bergum, Bruce O., and Lehr, Donald J. Effects of Authoritarianism on Vigilance Performance	75
Blaisdell, Francis J. Relationships between Carbon Chain Length and Avoidance Responses in Rats	284
Bolanovich, D. J. See West, Leonard J.	
Bottenberg, Robert A. See Madden, Joseph M.	
Bowers, David G. Self-Esteem and the Diffusion of Leadership Style	135
Briggs, George E. See Naylor, James C.	
Buck, Leslie. Auditory Perception of Position and Speed	177
Butterfield, Earl C., and Warren, Sue A. Prediction of Attendant Tenure	101
Campbell, David. Chance on SVIB: Dice or Men?	127
Campbell, David P., and Trockman, Rachel W. A Verification Scale for the Minnesota Vocational Interest Inventory	276
Campbell, John. See Dunnette, Marvin D.	
Carlson, Margaret E. See Hartman, Thomas F.	
Carp, Frances M., Vitola, Bart M., and McLanathan, Frank L. Human Relations Knowledge and Social Distance Set in Supervisors	78
Chylinski, Joanne. See Wright, Morgan W.	
Cline, Victor B., Richards, James M., Jr., and Needham, Walter E. Creativity Tests and Achievement in High School Science	184
Corcoran, D. W. J. Doubling the Rate of Signal Presentation in a Vigilance Task during Sleep Deprivation	412
Denton, J. C., and Prien, Erich P. Defining the Perceived Functions of Purchasing Personnel	332
Desrosiers, Wilfred. See Insel, Shepard A.	
Diers, Carol J. See Edwards, Allen L.	
Dunnette, Marvin D. A Modified Model for Test Validation and Selection Research	317
Dunnette, Marvin D. A Note on <i>The Criterion</i>	251
Dunnette, Marvin D., Campbell, John, and Jaastad, Kay. The Effect of Group Participation on Brainstorming Effectiveness for Two Industrial Samples	30
Edwards, Allen L. A Factor Analysis of Experimental Social Desirability and Response Set Scales ..	308
Edwards, Allen L., Walsh, James A., and Diers, Carol J. The Relationship between Social Desirability and Internal Consistency of Personality Scales	255
Ewen, Robert B. See Triandis, Harry C.	
Forehand, Garlie A. Assessments of Innovative Behavior: Partial Criteria for the Assessment of Executive Performance	206
Friedlander, Frank. Underlying Sources of Job Satisfaction	246
Gates, Stephen. See McKendry, James M.	
Geist, Harold. Work Satisfaction and Scores on a Picture Interest Inventory	369
Ghiselli, Edwin E. Moderating Effects and Differential Reliability and Validity	81
Gonyea, George G., and Lunneborg, Clifford E. A Factor Analytic Study of Perceived Occupational Similarity	166
Goodchilds, Jacqueline D. See Smith, Ewart E.	
Goodstein, Leonard D., and Schrader, William J. An Empirically Derived Managerial Key for the California Psychological Inventory	42
Graham, Warren R., and Johnson, Cecil D. An Experimental Comparison of Inventory Validity Obtained before and after Work Experience	72
Gustafson, Lawrence A., and Orne, Martin T. Effects of Heightened Motivation on the Detection of Deception	408
Harrell, Thomas W. See Lamouria, Lloyd H.	
Hartman, Thomas F., Morrison, Barbara A., and Carlson, Margaret E. Active Responding in Programmed Learning Materials	343
Hassler, Howard W., Myers, James H., and Seldin, Maurice. Payment History as a Predictor of Credit Risk	383
Hoff, Alvin L. See Bell, Forest O.	
Hoffman, L. Richard. See Maier, Norman R. F.	

Hoyt, Kenneth B. See Bell, Forest O.	
Insel, Shepard A., Schlesinger, Kurt, and Desrosiers, Wilfred. Dependency Responses to Televised Instruction.....	328
Jaastad, Kay. See Dunnette, Marvin D.	
Jacobson, Eugene, and Kossoff, Jerome. Self-Percept and Consumer Attitudes toward Small Cars...	241
Johnson, Cecil D. See Graham, Warren R.	
Johnson, Donald M. Reanalysis of Experimental Halo Effects.....	40
Johnson, Rossall J. Two Approaches to the Prediction of Group Responses.....	151
Jordan, Nehemiah. Allocation of Functions between Man and Machines in Automated Systems....	10
Keil, Rudolph C. See Kerr, Willard A.	
Kendall, L. M. See Smith, Patricia Cain.	
Kennedy, J. E., and Landesman, J. Series Effects in Motor Performance Studies.....	202
Kerr, Willard A., and Keil, Rudolph C. A Theory and Factory Experiment on the Time-Drag Concept of Boredom.....	7
Kinney, Jo Ann S. Night Vision Sensitivity during Prolonged Restriction from Sunlight.....	63
Kirchner, Wayne K., and Mousley, Nancy B. A Note on Job Performance: Differences between Respondent and Nonrespondent Salesmen to an Attitude Survey.....	223
Knapp, Robert R. Personality Correlates of Delinquency Rate in a Navy Sample.....	68
Kossoff, Jerome. See Jacobson, Eugene.	
Kuhlen, Raymond G. Needs, Perceived Need Satisfaction Opportunities, and Satisfaction with Occupation.....	50
Lamouria, Lloyd H., and Harrell, Thomas W. An Approach to an Objective Criterion for Research Managers.....	353
Landesman, J. See Kennedy, J. E.	
Latané, Bibb, and Arrowood, A. John. Emotional Arousal and Task Performance.....	324
Lehr, Donald J. See Bergum, Bruce O.	
Lepkowski, J. Richard. Development of a Forced-Choice Rating Scale for Engineer Evaluation....	87
Lunneborg, Clifford E. See Gonyea, George G.	
Lyman, John. See Ziedman, Kenneth.	
McKendry, James M., Snyder, Monroe B., and Gates, Stephen. Factors Affecting Perceptual Integration of Illustrated Material.....	293
Mackworth, Jane F. Effect of Reference Marks on the Detection of Signals on a Clock Face.....	190
McLanathan, Frank L. See Carp, Frances M.	
Madden, Joseph M., and Bottenberg, Robert A. Use of an All Possible Combination Solution of Certain Multiple Regression Problems.....	365
Maier, Norman R. F., and Hoffman, L. Richard. Seniority in Work Groups: A Right or an Honor?..	173
Marcus, Arthur. The Effect of Correct Response Location on the Difficulty Level of Multiple-Choice Questions.....	48
Mikesell, Eleanor Hall. See Triandis, Harry C.	
Miller, Irwin, and Minor, Frank J. Influence of Multiple-Choice Answer Form Design on Answer-Marking Performance.....	374
Minor, Frank J. See Miller, Irwin.	
Morrison, Barbara A. See Hartman, Thomas F.	
Mousley, Nancy B. See Kirchner, Wayne K.	
Mullins, Cecil J. Self-Confidence as a Response Set.....	156
Myers, James H. Predicting Credit Risk with a Numerical Scoring System.....	348
Myers, James H. See Hassler, Howard W.	
Naylor, James C., and Briggs, George E. Effect of Rehearsal of Temporal and Spatial Aspects on the Long-Term Retention of a Procedural Skill.....	120
Nayyar, Ravi M., and Simon, J. Richard. Effects of Magnification on a Subminiature Assembly Operation.....	190
Needham, Walter E. See Cline, Victor B.	
Norman, Warren T. Personality Measurement, Faking, and Detection: An Assessment Method for Use in Personnel Selection.....	225
Orne, Martin T. See Gustafson, Lawrence A.	
Osborn, William C., Sheldon, Richard W., and Baker, Robert A. Vigilance Performance under Conditions of Redundant and Nonredundant Signal Presentation.....	130
Park, James F., Jr., and Payne, M. Carr, Jr. Effects of Noise Level and Difficulty of Task in Performing Division.....	367
Patterson, C. H. See Rickard, Thomas E.	
Payne, M. Carr, Jr. See Park, James F., Jr.	
Porter, Lyman W. Job Attitudes in Management: II. Perceived Importance of Needs as a Function of Job Level.....	141

Porter, Lyman W. Job Attitudes in Management: III. Perceived Deficiencies in Need Fulfillment as a Function of Line versus Staff Type of Job.....	267
Porter, Lyman W. Job Attitudes in Management: IV. Perceived Deficiencies in Need Fulfillment as a Function of Size of Company.....	386
Poulton, E. C. Compensatory Tracking with Differentiating and Integrating Control Systems.....	398
Poulton, E. C. Pursuit Tracking with Differentiating and Integrating Control Systems.....	289
Pressey, Sidney L. Teaching Machine (and Learning Theory) Crisis.....	1
Prien, Erich P. Development of ■ Supervisor Position Description Questionnaire.....	10
Prien, Erich P. See Denton, J. C.	
Richards, James M., Jr. See Cline, Victor B.	
Rickard, Thomas E., Triandis, H. C., and Patterson, C. H. Indices of Employer Prejudice toward Disabled Applicants.....	52
Rigby, Marilyn K. See Severin, Francis T.	
Ronan, W. W. Work Group Attributes and Grievance Activity.....	38
Rowe, Patricia M. Individual Differences in Selection Decisions.....	304
Schlesinger, Kurt. See Insel, Shepard A.	
Schrader, William J. See Goodstein, Leonard D.	
Schultz, Duane P. Time, Awareness, and Order of Presentation in Opinion Change.....	280
Seldin, Maurice. See Hassler, Howard W.	
Severin, Francis T., and Rigby, Marilyn K. Influence of Digit Grouping on Memory for Telephone Numbers.....	117
Sheldon, Richard W. See Osborn, William C.	
Simon, J. Richard. See Nayyar, Ravi M.	
Sisler, George C. See Wright, Morgan W.	
Smith, Ewart E., and Goodchilds, Jacqueline D. Some Personality and Behavioral Factors Related to Birth Order.....	300
Smith, Patricia Cain, and Kendall, L. M. Retranslation of Expectations: An Approach to the Construction of Unambiguous Anchors for Rating Scales.....	149
Smith, Sidney L. Color Coding and Visual Separability in Information Displays.....	358
Snyder, Monroe B. See McKendry, James M.	
Spencer, Patricia A. See Andersen, L. Bryce.	
Strong, Edward K., Jr. Reworded versus New Interest Items.....	111
Tinker, Miles A. Influence of Simultaneous Variation in Size of Type, Width of Line, and Leading for Newspaper Type.....	380
Triandis, Harry C. Factors Affecting Employee Selection in Two Cultures.....	89
Triandis, Harry C., Bass, Alan R., Ewen, Robert B., and Mikesell, Eleanor Hall. Team Creativity as a Function of the Creativity of the Members.....	104
Triandis, H. C. See Rickard, Thomas E.	
Trockman, Rachel W. See Campbell, David P.	
Vitola, Bart M. See Carp, Frances M.	
Walsh, James A. See Edwards, Allen L.	
Warren, Sue A. See Butterfield, Earl C.	
West, Leonard J., and Bolanovich, D. J. Evaluation of Typewriting Proficiency Training: Preliminary Test Development.....	403
Whitlock, Gerald H. Application of the Psychophysical Law to Performance Evaluation.....	15
Wiener, Earl L. Knowledge of Results and Signal Rate in Monitoring: A Transfer of Training Approach.....	214
Wright, Morgan W., Sisler, George C., and Chylinski, Joanne. Personality Factors in the Selection of Civilians for Isolated Northern Stations.....	24
Ziedman, Kenneth, and Lyman, John. Effect of Variation in Task Complexity and Displayed Information on Operator Performance.....	260
Ziller, Robert C. Leader Assumed Dissimilarity as ■ Measure of Prejudicial Cognitive Style.....	339

Journal of Applied Psychology

KENNETH E. CLARK, Editor
University of Colorado

MARYGROVE COLLEGE LIBRARY
DETROIT, MICHIGAN
PLEASE DO NOT REMOVE

Table of Contents

Teaching Machine (and Learning Theory) Crisis: Sidney L. Pressey.....	1
A Theory and Factory Experiment on the Time-Drag Concept of Boredom: Willard A. Kerr and Rudolph C. Keil.....	7
Development of a Supervisor Position Description Questionnaire: Erich P. Prien.....	10
Application of the Psychophysical Law to Performance Evaluation: Gerald H. Whitlock.....	15
Personality Factors in the Selection of Civilians for Isolated Northern Stations: Morgan W. Wright, George C. Sisler, and Joanne Chylinski.....	24
The Effect of Group Participation on Brainstorming Effectiveness for Two Industrial Samples: Marvin D. Dunnette, John Campbell, and Kay Jaastad.....	30
Work Group Attributes and Grievance Activity: W. W. Ronan.....	38
An Empirically-Derived Managerial Key for the California Psychological Inventory: Leonard D. Goodstein and William J. Schrader.....	42
Reanalysis of Experimental Halo Effects: Donald M. Johnson.....	46
The Effect of Correct Response Location on the Difficulty Level of Multiple-Choice Questions: Arthur Marcus.....	48
Indices of Employer Prejudice toward Disabled Applicants: Thomas E. Rickard, H. C. Triandis, and C. H. Patterson.....	52
Needs, Perceived Need Satisfaction Opportunities, and Satisfaction with Occupation: Raymond G. Kuhlen.....	56
Night Vision Sensitivity during Prolonged Restriction from Sunlight: Jo Ann S. Kinney.....	65
Personality Correlates of Delinquency Rate in a Navy Sample: Robert R. Knapp.....	68
An Experimental Comparison of Inventory Validity Obtained Before and After Work Experi- ence: Warren R. Graham and Cecil D. Johnson.....	72
Effects of Authoritarianism on Vigilance Performance: Bruce O. Bergum and Donald J. Lehr	75
Human Relations Knowledge and Social Distance Set in Supervisors: Frances M. Carp, Bart M. Vitola, and Frank L. McLanathan.....	78

American Psychological Association

Consulting Editors

ARTHUR BRAYFIELD, *American Psychological Association*

GEORGE E. BRIGGS, *Ohio State University*

NORMAN FREDERIKSEN, *Educational Testing Service*

LEONARD D. GOODSTEIN, *University of Iowa*

EDWIN R. HENRY, *Standard Oil Company of New Jersey*

JOHN HOLLAND, *National Merit Scholarship Corporation*

CLIFFORD E. JURGENSEN, *Minneapolis Gas Company*

LAURENCE S. MCGAUGHRAN, *University of Houston*

QUINN MCNEMAR, *Stanford University*

HAROLD F. ROTHE, *Beloit Corporation*

THOMAS A. RYAN, *Cornell University*

CLARK L. WILSON, *Batten, Barton, Durstine, and Osborn, Incorporated*

This journal gives primary consideration to original investigations in any field of applied psychology except clinical psychology, although a descriptive or theoretical article may be accepted if it represents a special contribution in an applied field. Quantitative investigations of interest or value to psychologists working in the following broad fields will be considered: vocational and educational prognosis, diagnosis, and guidance at the secondary and college level; personnel research in business, industry, and government; engineering psychology; industrial working conditions; research on opinion and morale factors; job analysis and classification research; market and advertising research.

Manuscripts must be accompanied by an abstract of 100–120 words typed on a separate sheet of paper. The abstract should conform to the style of *Psychological Abstracts*. Detailed instructions for preparation of the abstracts appeared in the *American Psychologist* (1961, 16, 833), or they may be obtained from the Editor or from the APA Central Office.

Manuscripts should be addressed to the Editor:

Dr. Kenneth E. Clark
Office of the Dean
College of Arts and Sciences
University of Colorado
Boulder, Colorado

All manuscripts must be submitted in duplicate. Original figures are prepared for publication; duplicate figures may be photographs or pencil-drawn copies.

Manuscripts must conform to the style requirements described in the *Publication Manual of the American Psychological Association*.

The following policies govern reprinting of materials copyrighted in APA journals: (a) to require approval to reprint for tables and figures, and for text only if more than 500 words [total from one article] in length; (b) to continue approval to reprint, contingent on the author's approval, for articles reprinted in whole or in major part; (c) to negotiate where possible the dedication of royalties from commercial publishers to the American Psychological Foundation.

Journal of Applied Psychology

Published bimonthly by the
American Psychological Association
Prince and Lemon Sts., Lancaster, Pa.
and 1333 Sixteenth Street N.W.
Washington 6, D. C.

\$10.00 per volume

\$2.00 per issue

HELEN ORR
Managing Editor

ELIZABETH S. REED
Advertising Manager

VIRGINIA RICHARDS
Editorial Assistant

Subscriptions, orders, and business communications should be addressed to the American Psychological Association, 1333 Sixteenth St. N.W., Washington 6, D. C. Address changes must reach the subscription office by the tenth of the month to take effect the following month. Undelivered copies resulting from address changes will not be replaced; subscribers should notify the post office that they will guarantee second-class forwarding postage. Other claims for undelivered copies must be made within four months of publication.

Second class postage paid at Lancaster, Pennsylvania and at additional mailing places.

© 1963 by the American Psychological Association, Inc.

Journal of Applied Psychology

VOL. 47, No. 1

FEBRUARY 1963

TEACHING MACHINE (AND LEARNING THEORY) CRISIS¹

SIDNEY L. PRESSEY

University of Arizona

Current programmers are declared to be racing down a blind alley, misled by theory based on comparative studies whereas the important features of human learning of meaningful matter are largely unique to man. For such learning, there are indications that write-ins are less valuable than objective items, reinforcement not the significant process, programming destructive of cognitive structure. Objective autoinstruction adjunct to texts and other instructional matter is clearly most practicable, probably most learning-producing. Evaluations should compare this approach, as well as orthodox programmed learning, with conventional teaching. Research related to such endeavors should aid emergence of theory clarifying the distinctive features of human learning.

For several years now, all over the country, learning theorists have been programming books and other matter into numerous little "frames" each consisting of a very easy question or statement with space for writing a one or two word "constructed" response, to be verified by turning a page or turning up a "teaching machine" roll. One learned by responding (the theory was) and the more responding the more adequate the learning. In preparing each question the effort was not so much to contribute to a larger meaning as to assure that the student "emitted" the desired response, on the ground that he learned by making correct responses and an error would tend to recur. Multiple-choice questions are not used, because they involve the presentation of wrong alternatives, and also call merely for discrimination. All this has seemed plausible theoretically, and hopes have been high for extraordinary educational advances.

NOT GAIN BUT CONFUSION

Instead, evidence has been accumulating that the above hypotheses on which the programming was being based were, *for human learning of meaningful matter*, not so! Such

learners dealing with such materials may profit by seeing not only what a thing is but what it is not, may profit by mistakes, may learn to recall from learning to discriminate. Further, some half-dozen investigators have reported that as much may be learned in a given time simply by reading, as by reading *and* responding (Pressey, 1962; Silberman, 1962). In short, these theorists have independently discovered what educators have known about and been investigating for over 40 years—silent reading! Further, as programmed matter has been used over a period of time, it has been realized that for skimming for main ideas, for review—for any use except that initial go-through—the programmed book is almost impossible and the teaching-machine roll entirely so. Mostly, even for the first go-through, they are unsatisfactory, because most important matter to be learned has structure, which the programming destroys except the serial order, and most important learning is integrative and judgmental, so requires a looking about in what is being studied; for all such purposes a teaching machine seems about as hampering as a scanning device which required that one look at a picture only 1 square inch at a time, in a set order. Much seems very wrong about current attempts at autoinstruction.

¹ This paper was presented in modified form at the St. Louis meetings of the American Psychological Association, August 31, 1962.

A possible basic factor is suggested by Hilgard (1956) when he questions

the generalization from comparative studies that there are no differences, except quantitative ones, between the learning of lower mammals and man. . . . It is strange that the opposite point of view is not more often made explicit—that at the human level there have emerged capacities not approached by the lower animals, including other primates. . . . Language in man is perhaps the clearest of the emergents which carries with it a forward surge in what may be learned. . . . There are probably a number of different kinds of learning, following different laws. [Further, in man] the ceiling of ability itself may be modified by training. [Thus after acquiring] appropriate linguistic or mathematical tools [he can solve problems previously impossible] (pp. 460–461).

Surely that now taken-for-granted but really marvelous skill, silent assimilative reading, is such a tool. Also more important than often recognized are a variety of skills and strategies in learning usually grouped together as methods of study.

With Hilgard's position the writer would agree. He would say that the learning theorists have with notable vigor and consistency applied "generalizations from comparative studies" to problems of learning in school, and that the results have shown, more adequately than ever before, the unsatisfactoriness of those generalizations for that purpose. For a learner with reading-study skills, conventional textual matter orders and structures its contents in paragraphs and sections and chapters, exhibits that structure in headings and table of contents, makes all readily available in index with page headings and numbers. The learner thus has multiple aids to the development and structuring of his understanding. If need be he can, with a flick of the finger, move about in the material; he can skip the already known, turn back as a result of a later felt need, review selectively. As a way to present matter to be learned, the average textbook may not be best. But thousands of frames on a teaching-machine roll or strung through a programmed book would seem close to the worst. To make a very bad pun, the programmers have "framed" the textbook. Instead of trying to improve their programs, they might better consider very broadly how best to present matter for learning. The opinion is ventured that the best will be found closer to texts than to their programs.

But did not Socrates so teach the slave boy? The boy could not read. What about the often-cited skillful tutor? He assumed that the student had done some reading. However, both Socrates and the tutor did further learning by asking questions. The writer would contend that neither simply presented an idea and then reinforced it. Brownell's (1928) early research regarding primary school children's learning of arithmetic here seems relevant. Simply telling them that $2 \times 3 = 6$ did *not* bring about real learning of that number combination. These sturdy little empiricists had not merely to be *told*; they had to be *shown*, as by putting out two sets each of three pennies and demonstrating that they did indeed count to six. They had similarly to verify, and to differentiate, that $2 + 3$ was 5 and $3 - 2$ was only 1. As Piaget (1954) and others have described, children gradually develop a number system, also cognitive schema as of space, causality; and they do this not by so crude a rote process as the accretion of bit learnings stuck on by reinforcements, but by progressive processes of cognitive integration and clarification.

Moreover, such clarification is commonly by differentiation, and multiple-choice items involve just such processes. The three-choice question $2 \times 3 = 1, 5, \text{ or } 6$ differentiates the correct answer from answers got by wrongly subtracting or adding. In this one concise little item are thus packed three arithmetic processes and three number combinations, and study of the item might well involve all six issues, with autoinstructional dealing with the item clarifying of all. The point will be returned to.

But first a brief summary of the position so far. The past decade has seen an extraordinary "boom" in autoinstruction; most of this work has been dominated by concepts of operant conditioning deriving directly from animal experimentation and has become stylized in terms of initial presentation of tasks in numerous frames with immediate constructed response. Because thus so special in origin and nature, as well as yielding often question raising results, a basic critical review of current autoinstructional concepts seemed called for. Doubts have been raised as to whether human learning of meaningful

material can be adequately accounted for by animal based theory, programed matter is satisfactory for such learning, and reinforcement adequately accounts for the process (Gagné, 1962).

BUT WHERE FROM HERE?

When in doubt about such a theory-dominated situation, it is sometimes well to pull back and see whether a very practical analysis may helpfully reconstrue issues. If this be done, an obvious early question is this: what is the best way *initially* to present matter to be learned? The programmers have been cutting it into little pieces each responded to, but now recognize that one may learn from reading without responding. Then how big may the piece be? The writer has stressed that the bigger piece may have structure which should be made evident, and that first consideration as well as review or selective use may make it desirable that the learner can move about freely in the material. Perhaps it would be granted that a questioner who interrupted the reading of this paper should be asked to wait until it was all before him—that it would be then that the discussion could be most profitable. Surely it will be granted that the paper can best be understood if seen in print so that one can glance about and see headings; rather than if heard, when one cannot thus study—as one cannot study a teaching-machine roll. So the suggestion is: that the initial presentation might most often best be a very well organized and well written substantial statement much like a chapter in a good textbook! And the autoinstruction should follow and should be like a series of questions in a very good discussion of such a chapter.

Some “autopresentation” might be helpful: a teaching-machine roll might picture two groups each of three pennies and then six and so make clear to the child mind that 2×3 does make six. *After* his number system has been somewhat established, there may be automatized drill. The printed word “house” may be thus associated with a picture of one. Sundry sorts of detail-learning and of drill may be dealt with piecemeal. But mostly (the writer believes) initial presentation of what is to be learned will be in field trip,

demonstration or experiment, or most commonly a substantial unit like an incisive textbook chapter, *not* all mixed up with autoinstruction. The “autodiscussion” would follow, and its function would be (to paraphrase a statement in Ausubel’s 1961 review) to enhance the clarity and stability of cognitive structure by correcting misconceptions, and deferring the instruction of new matter until there had been such clarification and elucidation.

In difficult matter such as a science text or industrial or military training manual, bits of autoinstruction may be needed more frequently; each step in the solution of a difficult problem may need such autoelucidation. But the manual or text need not be fragmented into thousands of frames. Problems may be explicated in autoinstructional matter supplementary to the text; and there, or perhaps every 3 or 4 pages in the book, clusters of autoexplicating queries may keep check on understanding. But a book’s structured coherence and orderliness of presentation, and its convenience for overview, review, and reference, can be kept.

If the autoinstruction is thus to *follow* presentation of what is to be learned, then (like a good tutor or teacher) it will deal only with issues which need further clarification or emphasis. Such adjunct autoelucidation will *not* cover everything, may jump from one point to another or even back and forth. It will be very much shorter than present “programs,” which attempt both to present matter to be learned and autoinstruct about it in the same aggregate. Being so different, such supplemental autoinstruction might well be given a different name, as autoelucidation or explication.

But how would matter for adjunct autoinstruction or explication be selected? Experienced teachers would have many suggestions as to points needing special elucidation. They would be indicated in published research regarding pupils’ learning of and difficulties in spelling, arithmetic, algebra, composition, science, and history. Additional research, for development and trial of such elucidative material, would suggest more items and better ways of presenting them. Some could be cleared up by making the initial presentation

more lucid. But some students would still have difficulty with some items; perhaps those troubling 10% of the pupils or more would be dealt with in the adjunct autoinstruction.

The items should usually there appear (the writer is convinced) as multiple-choice questions with only such wrong alternatives as express common misunderstandings and a right answer notably clear. There is evidence that, contrary to theoretical inference, students do, after autoinstruction with such items, *less* often make the so-labeled mistakes, more often get things right, and transfer or generalize so that the gains appear on recall and yet other types of end tests (see for instance Jones, 1954; Lumsdaine & Glaser, 1960, pp. 52-93). Only half the students in a class may get such an item right on a pretest, but almost all of them do so on an end test a month later. In striking contrast, the perverse requirements of the orthodox programmer make any such effectiveness impossible: the item is initially supposed to be so easy that at least 95% pass it, errors cannot be identified as such because they must not be shown, and right statements are limited to such as the student can be maneuvered into hastily formulating himself. And orthodox improvement consists of making the items yet easier! In contrast, improvement of such an item as here urged would involve making wrong alternates clearer expressions of common misconceptions and the right more clearly right so that gains would be yet greater. In addition, the ease of checking objective items, with immediate indication of correctness (as by instant change of color of the check mark on a "chemo-card" or turn to next question on a key machine) makes possible going through many more items in a given time—so presumably more learning.

RANGE OF EVALUATIONS

But what of the argument that orthodox programs have been found greatly to save time, so that for instance a college course was finished in the first 2 months of a semester, or an industrial training course similarly shortened? Independent study plans have made possible marked reduction of time in class without any such programs (Baskin,

1960). The average class and the average business training session may be very time wasting and otherwise inefficient, and a number of alternatives may be shown to be better. In a college or secondary school course with several sections, it should be feasible to have one or more taught in conventional fashion, one or more use an orthodox program, a similar number try what the writer has called adjunct autoinstruction, another venture a planned independent study procedure, and outcomes on a carefully made final examination compared. If so made, such examinations can yield some analysis of outcomes: does one method or another bring more recall, transfer, application? Experiments of this type under the writer's direction have shown adjunct autoinstruction superior to conventional classes in all these respects.

These experiments also showed the adjunct materials very useful in planned independent study: in a room set aside for such use and having all the readings, laboratory material, and adjunct autoinstructional sheets available but looked after by an assistant, the students came in and worked when they wished, in small groups or individually, consulting the assistant when they so desired. All finished the 11-week course within 6 weeks. All did well on midterm and final examinations. But informal reports and interviews indicated yet other values, as gains in ability to work independently—though the students became better acquainted than in formal classes! The opportunity to save time was motivating. Several of these students took another course by independent study during the second half of the quarter.

More broadly, appraising experiments involving considerable numbers of students with different instructors over considerable periods of time—preferably a whole school or business training course—have yet other values. Methods have to be tolerable in long continued and routine, not simply brief and special, use. In the work just described, the best all-purpose "teaching machine" was judged to be a 3×5 chemo-card having 30 lines each of four squares: on this answer card the student checked his choice of answer to each of 30 four-choice questions on a teach-test sheet, using a special red ink which

instantly turned black when he marked in the right answer-box (because of an invisible chemical printed there). The student kept trying on each question until this color-change feedback told him he had the correct answer. For remedial review he had only to note where his red marks were, the sum of them was his error-score; the instructor had only to note where he saw most red on the cards for a given day to see where some corrective discussion might be desirable, and for both him and the students the cards were a compact easily-filed record.² In the writer's adjunct autoinstructional procedure, everything except the cards could be used over and over again, easily returned to again as for review. For long-continuing flexible use and re-use, it seemed apparent that a text or business manual plus perhaps 50 adjunct autoinstructional sheets (and some chemo-cards) was far more practicable than that manual or text cut up into 3,000 frames on a teaching-machine roll (with the machines) or strung through a programed book.

RESUME AND RECOMMENDATIONS

Teaching machines and programed materials are now being used all over the country in schools and colleges and in industrial and military training. Manufacture and sale of such products are a major enterprise of many publishers and equipment makers. Ambitious young people are embarking on careers in such work. The whole subject has become an accepted topic of everyday talk. However, there is disturbing evidence that current autoinstruction is *not* up to the claims made for it, that the current "boom" might be followed by a "bust" unfortunate for those involved—and for psychology. This paper is first of all a plea that to guard against such a danger the whole situation be soon given close critical

² Yet more convenient autoinstructional cards are possible. Instead of a pen with special ink, only a pencil may be needed; a mark with it, or a stroke of its eraser, breaks through an overprint to reveal a "c" underneath when the right answer is found. For 30-item 3-choice teach tests, a device little larger than a stop watch, and less complicated, may both teach and keep score. An apparatus little larger than an electric desk clock may both teach and provide selective review.

inspection, and not merely to assure (as is now being attempted) that programs are good; but critically to consider whether the whole current concept of programing may be at fault, and an almost totally different approach than now orthodox to all ideas about autoinstruction be called for.

The archvillain, leading so many people astray, is declared to be learning theory! No less a charge is made than that the whole trend of American research and theory as regards learning has been based on a false premise—that the important features of human learning are to be found in animals. Instead, the all-important fact is that human has transcended animal learning.³ Language, number, such skills as silent reading, make possible facilitations of learning, and kinds of learning, impossible even for the apes. Autoinstruction should enhance such potentials. Instead, current animal derived procedures in autoinstruction destroy meaningful structure to present fragments serially in programs, and replace processes of cognitive clarification with largely rote reinforcings of bit learnings.

An "adjunct autoinstruction" is urged which keeps, makes use of, and enhances meaningful structure, the autoinstruction serving to clarify and extend meaningfulness. Texts, manuals, laboratory exercises, instructional moving pictures and television would be kept (though often improved), and the autoinstruction would aid in their use and increase their value. The materials would be perhaps only a tenth as bulky as present programs; and being objective, their use

³ For this conclusion there is no less evidence than the whole history of civilization! Basically more significant than Skinner's brilliant research regarding animal learning may well be the almost forgotten finding of Kellogg and of Cathy Hayes that even if an ape be raised in a home like a child, it can never learn to talk. Far more remarkable than Skinner's pigeons playing ping pong is the average human scanning a newspaper—glancing about to find matter of interest to him, judging, generalizing, reconstruing, all in silent reading without overt respondings or reinforcings. Most remarkable of all is it to see learning theorists, hypnotized by the plausibilities of a neat theory, trying to teach that human as if he were a pigeon—confining his glance to the rigid slow serial peep show viewing of innumerable "frames" each demanding that he respond and be reinforced.

could be greatly facilitated by automating devices.

Evaluations should not merely (as is now projected) compare the merits of various "orthodox" programs. Those should be compared with such adjunct autoinstructional materials as here advocated. Adaptability should be compared for use with other media as books and movies and other methods as guided independent study. Convenience and cost for continuing general use should be hard-headedly appraised. The prediction is ventured that in all respects adjunct autoinstruction will be found far superior: time and work saving will be great yet more will be accomplished—courses often completed in half the usual time, years saved but nevertheless more accomplished in school and college, industrial and military training tasks reduced perhaps a third in length and all with great time and trouble saved instructional staffs. Then at long last the "industrial revolution" in education may come about which the writer predicted (Pressey, 1932) just 30 years ago. Further, somewhat as the practical testing movement from the first world war on greatly stimulated and aided research and theorizing regarding abilities, so autoinstruction may get research on learning out from under its long dominance by com-

parative psychology and confinement in the laboratory and evolve vigorous new theory.

REFERENCES

- AUSUBEL, D. P., & FITZGERALD, D. Meaningful learning and retention: Intrapersonal and cognitive variables, *Rev. educ. Res.*, 1961, **31**, 500-510.
- BASKIN, S. *Quest for quality: Some models and means*. Washington, D. C.: United States Department of Health, Education, and Welfare, 1960.
- BROWNELL, W. A. The development of children's number ideas in the primary grades. *Suppl. educ. Monogr.*, 1928, No. 35.
- GAGNÉ, R. M. Military training and principles of learning. *Amer. Psychologist*, 1962, **17**, 83-91.
- HILGARD, E. R. *Theories of learning*. (2nd ed.) New York: Appleton-Century-Crofts, 1956.
- JONES, R. S. Integration of instructional and self-scoring measuring devices. *Abstr. doct. Dissert., O. State U.*, 1954, **65**, 157-165.
- LUMSDAINE, A. A., & GLASER, R. *Teaching machines and programmed learning*. Washington: National Education Association, 1960.
- PIAGET, J. *The construction of reality in the child*. New York: Basic Books, 1954.
- PRESSEY, S. L. A third and fourth contribution toward the coming "industrial revolution" in education. *Sch. Soc.*, 1932, **36**, 668-672.
- PRESSEY, S. L. Basic unresolved teaching machine problems. *Theory Pract.*, 1962, **1**, 30-37.
- SILBERMAN, H. F. Self-instructional devices and programmed materials. *Rev. educ. Res.*, 1962, **32**, 179-193.

(Early publication received October 15, 1962)

A THEORY AND FACTORY EXPERIMENT ON THE TIME-DRAG CONCEPT OF BOREDOM

WILLARD A. KERR

AND

RUDOLPH C. KEIL

Illinois Institute of Technology, Chicago

Wolber Duplicator and Supply Company

47 (81%) of 53 personnel of a small plant guessed how much and in which direction the "erroneous" (but actually correct) company clocks were in error, in order to test hypothesis that, contrary to popular notion, time drag is estimated as greater in the variety-type than in the repetitive-type jobs. Significant correlations ($p < .05$) of time drag with variety-type (.43) rather than monotony-type, and with long-cycle (.27) rather than short-cycle jobs resulted. Time drag actually was greater in variety-type and in long-cycle jobs. The hypothesis that unbroken time is perceived analogously to unbroken visual space seems tenable.

It is popular belief that when time seems characteristically to "drag" (pass slowly) in a job, the job is interpreted by the worker as boring. Further, the common belief that the short-cycle (repetitive) job is conducive to boredom *because* it elicits a feeling of time drag has passed virtually unchallenged in the literature (Karn & Gilmer, 1952, pp. 239-249). Yet extrapolation by analogy of an established principle in visual space perception suggests a theory which abruptly contradicts these popular ideas.

As Scott (1931) clearly demonstrated, the viewing of optimally interrupted space as contrasted with solid or uninterrupted space results in a perception that the former (actually equal) space is significantly the greater. If the breaking up of visual *space* by dots, lines, or objects causes it to seem greater, is it not plausible that the breaking up of *time* by significant psychological events will cause time to seem greater? The temporal perception theory here formulated then is simply that when a subject's perceptual time is interrupted by attention-demanding events, such time will be judged as *greater* than if equivalent objective time is not so interrupted. From this it follows, if the theory is valid, that workers on the repetitive (short-cycle, overlearned) jobs, working in the relative absence of significant psychological time markers, will find time to pass *faster*, not slower, than it does for personnel in the events-studded nonrepetitive jobs. A circumstantial clincher to this argument is that time goes fastest of all in that total absence of

significant event markers known as "sound" sleep.

The theory just stated permits us to postulate a perceptual time-speed continuum ranging from utterly sound sleep through the overlearned short-cycle tasks that one can do successfully in a semistupor through the overlearned moderate-cycle tasks to the more unpredictable-events sequence jobs. In the latter jobs, these paradoxical events transpire: (a) time will be judged objectively to pass more slowly than in the short-cycle jobs; (b) individuals will be unaware of point *a* just stated and will, in fact, maintain the fiction that time passes more slowly in the short-cycle jobs. Since the last point is a common and obvious observation, the pilot experiment which follows deals primarily with job-cycle length and objective judgment of time lapse plus certain other variables.

EXPERIMENTAL DESIGN

Subjects. Forty-seven (81%) of the 53 personnel of the Wolber Manufacturing Company cooperated in this research. Age range was 21-60, the mode being 47. Forty-two were male, five female. All personnel were under hourly rate pay system. Three employees were literate only in German and six only in Polish.

Procedure. At a noontime break the plant manager announced to the assembled personnel that:

This afternoon I.I.T. will give a little research survey. You may cooperate or not, as you wish. The management will not handle the data and will be informed only of the general results.

Between 2:00 and 2:30 P.M. on the same afternoon, an employee conspicuously "worked" on the

TABLE 1
INTERCORRELATIONS AMONG EIGHT VARIABLES RELATIVE TO TIME-DRAG
PERCEPTIONS OF 47 PERSONNEL

	2	3	4	5	6	7
1. Time drag (overestimation)	-.43 ^a	.03	-.46 ^a	.21	.27 ^a	-.42 ^a
2. Monotonous-type job		-.33	.36	.03	-.82	.30
3. Interest (versus boredom) in job			-.04	.23	.09	.25
4. Service on present task				.14	-.61 ^a	.44 ^a
5. Present task is less repetitive than previous one					-.26	-.13
6. Job cycle length						.36
7. Accuracy of time estimate						

^a Coefficients which are acceptable at the 95% level of probability.

shop clocks. Immediately afterward a brief trilingual (English, German, Polish) questionnaire was distributed. It advised personnel to:

Please answer the following questions but DO NOT sign your name to this paper. This is a project to determine how accurately persons at work can estimate time. DO NOT LOOK AT YOUR WATCH. At this moment the company clocks are either slower or faster *than they usually are*. Without talking to anyone, please answer the following: I think the clocks are about_____ minutes (check one) fast_____, slow_____. At present I am working on_____. Before this I was working on_____. That was about_____days ago. I feel that my work is: _____very interesting; _____interesting; _____boring; _____very boring. Thank you very much for your cooperation.

In addition to the data obtained on this questionnaire, two other variables were included in the study: (a) length of basic time cycle for the total sequence of operations involved in each job, (b) subjective classification of each job as "monotony-type" or "variety-type." All data were intercorrelated.

RESULTS

As indicated in Table 1, time drag was greater in variety-type than in monotony-type jobs ($r = .43$); and it was greater in long-cycle than in short-cycle jobs ($r = .27$), supportive of the stated theory and contrary to popular notions. Also, in this study, absolute error in time estimate is related significantly to the time drag ($r = .42$); time not only passed more slowly but also less accurately for the personnel who exhibited time drag. Experience on a job is related ($r = .44$) to accuracy of time estimate.

In further confirmation of the theory, the

employee's own anonymous statement of interest-boredom in the job is unrelated to the time drag ($r = .03$), suggesting popular semantic confusion in verbalizing about the psychic income from work experience. Incidentally, the significant relationship between job-cycle length and time drag remains unchanged (partial $r = .27$) when expressed interest-boredom is held constant.

Length of service on a particular job is significantly related to time drag ($r = -.46$) and supports the hypothesis that experience makes the job less challenging, leaving the mind free for "time-unmarked" activity and, hence, decreased time drag. Time drag is not significantly less ($r = .21$) when the present job is more monotonous than the previous one, but the relationship is in the direction anticipated, namely, that time drag is greater when the present job is less repetitive than the previous job.

INTERPRETATIONS AND CONCLUSIONS

The above experimental results are not inconsistent with the early reports of both Kries (1923) and Sturt (1925) that time drag is determined by the amount of mental content experienced and may be independent of the job performed. The results do contradict, however, the observations given by Wyatt (1929):

the bored individual . . . is inclined to over-estimate the duration of time . . . when an activity is accompanied by interest . . . time in consequence appears to pass comparatively quickly . . . in repetitive

processes devoid of interest . . . the day seems . . . overestimated.

The popular views, well stated and apparently endorsed by Wyatt, collide sharply with research evidence.

Interest in the job, which presumably would also mean ego-involvement, is not a significant predictor of time drag, despite the almost universal tendency of personnel to say "Oh yes, when I'm doing work that is interesting, time seems to fly." Perhaps what people really mean is that when time is expended with *deeply satisfying* results it seems to fly. In other words, a more purely hedonistic approach to the verbalizing may help clear up

the semantic confusion; this, of course, needs research verification.

REFERENCES

- KARN, H. W., & GILMER, B. H. Readings in industrial and business psychology. New York: McGraw-Hill, 1952.
- KRIES, V. Allgemeine Sinnesphysiologie. Leipzig: F. C. W. Vogel, 1923.
- SCOTT, W. D. The psychology of advertising. New York: Dodd-Mead, 1931.
- STURT, M. The psychology of time. London: Paul, Trench, Trubner, 1925.
- WYATT, S. Boredom in industry. *Personnel J.*, 1929, 8, 161-171.

(Received October 16, 1961)

DEVELOPMENT OF A SUPERVISOR POSITION DESCRIPTION QUESTIONNAIRE¹

ERICH P. PRIEN

Psychological Research Services, Western Reserve University

A study of the job duties of factory foremen. A Supervisor Position Description Questionnaire (SPDQ) was developed and administered to 24 factory foremen and the corresponding supervising executives. An inverse interbattery factor analysis was performed. 7 factors were obtained and titled: Manufacturing Process Supervision; Manufacturing Process Administration; Employee Supervision; Manpower Coordination and Administration; Employee Contact and Communication; Work Organization, Planning, and Preparation; and Union Management Relations. A 2nd factor analysis of SPDQ scores yielded 2 factors titled: Manufacturing Operations Management and Administration, and Manpower Management and Utilization. The factors are compared to those obtained by Hemphill (1961) and to the results of leadership studies.

A recent development related to the identification and measurement of management-position functions is the study by Hemphill (1961). Hemphill developed and analyzed a position description questionnaire to identify the functions of executives common to different positions in different organizations. The method *does not* identify or measure job functions which are unique to a single job or to one company. However, when the procedures and analysis are repeated using data from a single company, the stable job functions are identified, limited only by the scope of the questionnaire.

The purpose of this study was to develop criterion dimensions for first-line supervisory positions.

Design and Procedure

The basic design used in this study is an inverse, interbattery factor analysis of the responses to the Supervisor Position Description Questionnaire (SPDQ). The factor analysis procedure was developed by Tucker (1958).

Development of Questionnaire. Prior to writing the questionnaire statements, a listing was made of the general functions usually associated with first-line supervisory positions. Statements were written by several individuals who were qualified both as

job analysts and as item-writers. Additional statements were written which did not fall into any one of the outlined areas or which were unique to the positions being studied. Finally, each statement was reviewed for clarity and the format selected for the questionnaire.

Administration of Questionnaire. A small-group procedure was used for administration of the SPDQ. It was decided that the benefits of conducting small group meetings outweighed the disadvantages. However, each meeting was as nearly as possible a replication of the first meeting.

For this study, the sample consisted of 24 first-line supervisors in one company installation and 6 supervising executives.

RESULTS

The interbattery analysis results are in terms of clusters of jobs which are identified by factor loadings. These appear as Tables 1 and 2.

At this point in the analysis, the factors are not yet psychologically meaningful. Additional computations yield statement weights identifying the precise nature of the similarity of each cluster of jobs.

The seven factors derived in this study were defined by a cluster of job titles and more specifically by a cluster of questionnaire statements.

Factor 1: Manufacturing Process Supervision. The foremen performing this function identified by Factor 1 are concerned with process operations on a day-to-day basis. The activity is perpetual and involves rather constant attention to maintain the basic manufacturing operation. In the performance

¹ The research upon which this article is based was originally presented at the Executive Study Conference, Educational Testing Service, June 1961. The proceedings of this conference are published by Educational Testing Service. I wish to acknowledge the collaboration of Edgar F. Huse of the Raytheon Corporation in the analysis of the research data and the definition of the factors.

of this function, the foreman is not involved with other personnel or problems for which action can be postponed. This function usually requires immediate appraisal and action and involves an emergency characteristic to some degree. The scores on this factor were higher for line positions than for maintenance positions.

Examples of statements used in defining this factor are: adjust process operations to facilitate maintenance work, initiate start-up and shut-down procedures, and participate in adjusting process schedule to allow for maintenance work.

Factor 2: Manufacturing Process Administration. Performance of this function involves inspection and surveillance of manufacturing operations. The emphasis is on maintaining the operation on a continuous and efficient basis. The major concern is the manufacturing operation as affected by things other than equipment and raw materials. Although the elements have an administrative connotation,

TABLE 1
ROTATED FACTOR LOADINGS: BATTERY I
EXECUTIVES

	A	B	C	D	E	F	G
1.	.00	.34	.38	.44	-.11	.01	.14
2.	-.11	.24	.28	-.03	-.09	-.05	-.09
3.	.01	.38	.22	.07	.04	-.11	-.13
4.	-.07	.30	.18	.15	.03	-.09	-.21
5.	-.03	.31	-.43	.02	-.11	-.08	.04
6.	-.03	.31	-.43	.02	-.11	-.08	.04
7.	-.02	.31	-.43	.03	-.12	-.07	.04
8.	-.07	.30	-.03	.18	-.24	.08	-.02
9.	-.02	.30	.08	-.37	-.15	-.26	.01
10.	-.01	.28	.21	-.53	.04	-.42	.11
11.	-.27	-.07	.10	.09	-.09	.01	-.25
12.	-.30	-.07	.01	.04	-.23	-.16	.06
13.	-.27	-.02	.02	.24	-.16	-.09	-.05
14.	-.28	-.07	-.04	.24	-.22	-.44	.67
15.	-.30	-.06	-.09	-.02	-.10	-.08	-.09
16.	-.28	-.07	-.14	-.15	.03	.01	-.20
17.	-.28	-.08	.12	-.22	-.15	-.09	-.10
18.	-.30	-.08	-.02	.08	-.10	-.04	-.15
19.	-.25	.03	.07	-.12	-.57	.01	.10
20.	-.25	.03	.01	-.09	-.47	.03	.00
21.	-.21	.05	.02	-.07	-.41	.08	-.07
22.	-.22	.04	.10	-.11	-.57	.06	.15
23.	-.22	.04	.11	.01	-.53	.10	.10
24.	-.28	.04	-.15	-.04	-.49	.13	-.16

Note.—Significant loadings are bold faced.

TABLE 2
ROTATED FACTOR LOADINGS:
BATTERY 2 FOREMEN

	A	B	C	D	E	F	G
1.	-.15	.54	.06	.15	.11	-.09	-.14
2.	-.17	.43	.06	-.09	-.13	.05	.07
3.	-.12	.49	.13	-.09	-.07	-.04	.13
4.	-.22	.57	.13	-.06	.25	-.05	-.22
5.	-.02	.69	.08	.04	.00	-.03	.08
6.	-.27	.65	-.04	-.03	.07	-.08	.10
7.	-.12	.70	-.26	.14	-.02	.10	-.02
8.	-.16	.70	.09	.22	-.16	.19	.02
9.	-.35	.70	.13	-.31	-.12	-.01	.10
10.	-.25	.34	-.07	-.10	.33	-.20	-.06
11.	-.78	-.08	.11	.17	.13	.02	.01
12.	-.63	.09	.05	-.18	-.07	.33	-.43
13.	-.73	.20	-.08	.06	.08	.01	.03
14.	-.71	.12	-.07	-.06	-.15	.09	.11
15.	-.81	.12	-.04	.12	.16	-.08	.14
16.	-.71	.22	-.23	-.08	-.17	.08	.20
17.	-.75	.11	.14	.07	-.06	.13	.02
18.	-.78	-.20	.29	.00	.05	-.10	.18
19.	-.59	.16	.02	-.13	.14	-.07	.00
20.	-.64	.21	-.06	.06	.04	.08	-.03
21.	-.48	.08	-.14	.01	.11	-.01	-.02
22.	-.72	.20	-.04	.04	-.00	.15	-.05
23.	-.38	-.12	-.19	-.20	.17	-.12	.00
24.	-.67	.25	-.02	.09	-.50	-.50	-.26

Note.—Significant loadings are bold faced.

personnel and personnel problems are not involved. The emergency characteristics of Factor 1 are not involved in Factor 2 even though the ultimate objective is similar—that of maintaining the manufacturing operation. (The process jobs as a group scored higher on this factor than any maintenance jobs.)

Examples of statements used in defining this factor are: periodically check quality control reports, report and prepare requests for maintenance and repair work, and observe subordinates' performance.

The essential character of Factors 1 and 2 is that both are oriented to the manufacturing process operation. Both factors require knowledge of the manufacturing process and elements which affect efficiency and continuity of operation. In that the sources of activity of both factors are similar and in some respects identical, it is logical that the responsibility for both is vested in one individual.

Factor 3: Employee Supervision. The individuals performing this function are con-

cerned with face-to-face relationships of a personal nature, but still within and with reference to the organization. The nature of the function is active; the individual participates in interpersonal relations, he does things either according to previously laid plans or on the basis of his judgment at that time. Although this is an active function, the planning, administration, and policy making relevant to the actions are not included. These aspects are separate and independent.

Examples of statements used in defining this factor are: explain and discuss problems with staff personnel, explain personnel policies, procedures, and benefit plans to subordinates, and schedule job rotations for employees.

Factor 4: Manpower Coordination and Administration. Individuals performing this function plan, prepare for, and administer with specific reference to the use of employees in the administering operation. This is the administrative function necessary to make effective use of employees in the manufacturing operation. This is a passive function as compared to Function 3, although this activity leads to or follows an action. The function, even though employee oriented, is not in the nature of personnel administration.

Examples of statements used in defining this factor are: prepare maintenance-completed reports, delegate authority to group leaders to plan and carry out a task, and contact staff personnel.

Factor 5: Employee Contact and Communications. Individuals performing this function have frequent face-to-face contact with employees. The contact may center about the job, but more likely a nonjob topic. In character, this function is similar to that of Function 1, in that it is active. The individual does things, he participates or executes plans, policies, or decisions. However, the activities are employee centered even though the basis for an act could be impersonal, the result of executive decision, or conditions beyond the control of the employee or the foreman. Communications enter into the function to a considerable degree and generally are with subordinate employees or with reference to subordinate employees.

Examples of statements used in defining

this factor are: provide task instructions for subordinates, discuss quality control with subordinates, and review applicant's personal history and work record.

*Factor 6: Work Organization, Planning, and Preparation.*² Individuals who perform this function are concerned with making plans and preparations for employee utilization. The work directly involves material, equipment, and the environment; but the objective is to provide the necessities for manpower utilization. While some personal contact is involved, this is incidental to the main objective and most of the work is individual.

Examples of statements used in defining this factor are: prepare yearly vacation schedules, attend refinery operation-coordination meetings, and provide for movement of equipment and materials to and from work site.

Factor 7: Union-Management Relations. Individuals performing this function work within and participate in the creation of union-management relations. The function is fairly active, involves contact, and is employee-and-manpower-utilization oriented. The individual does things which have immediate as well as long-term consequences which are related to employees, but do not involve employees directly in the situation. As in Function 5, manufacturing operations, as such, are not involved.

While the concept of this function is quite clear, executives place much more importance on it than do the foremen.

Examples of statements used in defining this factor are: discuss formal first-step grievances with union committeemen, discuss misunderstandings with union committeemen, and initiate disciplinary action.

The functions defined by Factors 6 and 7 are similar in that both are concerned with elements affecting employee utilization, but do not involve employees directly. While some activity is involved, the functions to a much greater degree involve planning, preparation, organization, and administration.

² There was a greater amount of disagreement between foremen and executives on this factor than on any other. Many items with a high positive weight for foremen had high negative weights for the executive battery.

Second Factor Analysis

A second direct factor analysis was performed to facilitate the interpretation of the factors obtained in the first analysis. The results of this analysis are directly interpretable in terms of the first factor definitions. Correlations were computed using the factor scores for the 48 questionnaires. A direct Centroid Factor Analysis was performed. Two factors were obtained and rotated graphically to facilitate interpretation. The rotated factor loadings appear in Table 3.

Factor I appears to be a job or work oriented factor. Quite prominent is the responsibility for Work Supervision and Administration. The low negative, but significant loading of the employee oriented element supports the job oriented interpretation. This factor is named Manufacturing Operations Management and Administration.

Factor II appears to involve an employee oriented concept. The action element of the concept reflects face-to-face contact and participation with employees. Supporting this are significant negative loadings on administrative factors involving employee utilization. This concept is named Manpower Management and Utilization.

It seems, from this analysis, that the role of the first-line foremen in this manufacturing operation involves two basic functions; namely, dealing with people and production. In turn, each of the major dimensions consists of more specific functions.

The factors as derived appear to make up a general structure which is both convenient and logical. The composite foreman position involves both the physical organization (production) and the human organization (employees) directly.

The structure is not presented as the ultimate or as the only scheme for interpreting the obtained results. However, this interpretation is consistent with current thinking in the field of management philosophy and organization theory. Within the general structure presented in this study, it is conceivable to maintain manufacturing operations in the absence of any attention or emphasis on the purely employee oriented functions. Herzberg (1959), Argyris (1960), and others are quite convinced that this is the current practice in industry today, and further, that while failure to attend to the employee oriented aspects of the business may not have an immediate negative impact on profits, the long-range implications can be serious.

The results obtained by Katz and his associates (1950, 1951), and the leadership studies performed at Ohio State University, Fleishman (1955), support the view that supervision can emphasize either the employee oriented or human-relations functions of supervision. The results of this study, particularly the second factor analysis, are consistent in that obtained factors describe a personal, employee oriented versus an impersonal, job oriented scheme.

Another possible structure for interpreting the factors obtained in this study involves productivity/profitability, and custodial functions. In this scheme—referring to the first factor analysis—Factors 1, 2, and 6 are interpreted as profit oriented functions while Factors 3, 4, 5, and 7 are viewed as custodial functions. While this is also a meaningful structure, it is not as “clean” nor as psychologically meaningful.

Comparison of the results obtained in this study with those obtained by Hemphill reveals considerable similarity. Dimensions identified by Hemphill which do not appear in this study are: Technical Markets and Products; Human, Community, and Social Affairs; Business Reputation; and Personal Demands. That these functions were not identified is not inconsistent when the sample characteristics are considered. Only one level of the hierarchy of positions is sampled and only in the manufacturing operations portion of the business.

TABLE 3
ROTATED FACTOR LOADINGS
FOR SECOND ANALYSIS

Factor	A	B	C	D	E	F	G
I	78	90	-21	-54	-07	15	22
II	00	-12	03	-14	71	-27	-63

Note.— $N = 48$.

TABLE 4
CORRELATION BETWEEN FACTOR SCORES AND FACTOR
LOADINGS OVER ALL QUESTIONNAIRES

	Factor						
	A	B	C	D	E	F	G
Correlation	.62	.76	.87	.58	.48	.62	.46

On the other hand, the functions identified in the current study appear to be well-covered by the remaining five dimensions identified by Hemphill. The seven dimensions identified in this study appear, as expected, to be more specific.

The procedure used in identifying the questionnaire items scored for each factor is designed to replace the initial job factor loading. The degree to which the factor scores approximate or predict the job factor loadings are represented by the correlations in Table 4.

The results of this study provide support for the development of procedures to describe position functions of an intangible nature.

REFERENCES

ARGYRIS, C. *Understanding organization behavior*. Homewood, Ill.: Dorsey, 1960.

FLEISHMAN, E. A., HARRIS, E. F., & BURT, H. E. Leadership and supervision in industry. *Ohio St. U. Stud. Bur. Educ. Res. Monogr.*, 1955, No. 33.

HEMPHILL, J. K. Dimensions of executive positions. *Ohio St. U. Stud. Bur. Educ. Res. Monogr.*, 1961, No. 98.

HERZBERG, F., MAUSNER, B., & SNYDERMAN, BARBARA. *The motivation to work*. New York: Wiley, 1959.

KATZ, D., MACCOBY, N., GURIN, G., & FLOOR, L. *Productivity, supervision, and morale among railroad workers*. Ann Arbor, Mich.: University of Michigan, Survey Research Center, 1951.

KATZ, D., MACCOBY, N., & MORSE, N. *Productivity, supervision, and morale in an office situation*. Ann Arbor, Mich.: University of Michigan, Survey Research Center, 1950.

TUCKER, L. R. An interbattery method of factor analysis. *Psychometrika*, 1958, 23, 111-136.

(Received October 23, 1961)

APPLICATION OF THE PSYCHOPHYSICAL LAW TO PERFORMANCE EVALUATION

GERALD H. WHITLOCK

University of Tennessee

In the area of job performance, it was hypothesized that evaluations, e.g., "poor" to "excellent," were based on the observation of "performance specimens" where a performance specimen is defined as "an incident of relevant performance which at the time of observation was classed as uncommonly effective or uncommonly ineffective." It was further hypothesized that the psychophysical law ($y = kx^n$) would describe the relationship between the number of specimens observed (x) and resulting evaluations (y) of performance. Finally, it was hypothesized that Steven's criterion for prothetic continua would be satisfied—concave downward curve when ratio estimation scale values for sets of performance specimens are plotted against corresponding category scale values. Using simulated performance ratings as well as actual performance evaluations (performance ratings of apprentices, professors, supervisors, and executives) the above hypotheses appear to have been verified.

Psychophysics is concerned with relationships between physical and psychological magnitudes. This paper is concerned with the relationship between observation and evaluation of performance. Stevens has presented evidence that a power function describes the psychophysical relation for at least 20 continua with the consequence that to him the psychophysical law can be stated simply as, "Equal ratios in stimulus produce equal ratios in response." This paper purports to show that a power function also describes the relation between observation of performance and evaluation of performance for four different types of endeavors.

The model underlying the theory of performance evaluation here presented has the following assumptions: (a) evaluation of performance is a response to a set of observations of performance, (b) observations which are associated with evaluation of performance are the observations of performance specimens, and (c) observations of specimens can be remembered over a reasonable rating period (e.g., 6 months) and reported with sufficient accuracy at the end of the period. A performance specimen is an incident of relevant performance which is uncommonly effective or uncommonly ineffective. The set of performance specimens for a particular performance area is greater than the set of critical incidents for that area, but the entire set of critical incidents

is included in the set of specimens. This is to say that the set of performance specimens will include performance which is nearer than critical incidents to the mean of the distribution of the total set of performances in an area when all of the performances are ordered on a continuum of effectiveness.

Performance specimens for a particular area of performance may be collected by obtaining from a sample of experts in that area incidents of performance observed by them which at the time of observation were classed as uncommonly effective or uncommonly ineffective.

It follows that four possible conditions exist with respect to the set of performances for an individual over some observation period.

1. His set of performances may contain no performance specimens. That is, there may have been no performances during the observation period which were nonaverage.

2. The set of his performances may contain one or more performance specimens all of which were effective performances.

3. The set of his performances may contain one or more performance specimens all of which were ineffective.

4. The set of his performances may contain performance specimens some of which were effective and some of which were ineffective.

It was hypothesized that a lawful relation obtained between the number of performance specimens observed for an individual during an observation period and the overall evaluation of that performance. The general nature of this relationship was first assumed to be that increments in evaluation grew as the positive difference between the number of effective and ineffective specimens observed. It was later hypothesized that Fechner's law would describe the relation between perception and evaluation of performance (Beard, 1960). However, neither the linear nor the logarithmic function was found to be appropriate. The following experiment to test the latter was conducted last fall (Hedge, 1960).

Performance specimens in the area of executive performance were collected by the writer several years ago from approximately 250 executives in a chemical plant. Certain of these specimens were used to prepare a mock performance evaluation check list, consisting of 16 specimens each of effective and ineffective performance. These forms were prechecked with all combinations of 1, 2, 4, 8, and 16 specimens of effective and ineffective performance.

In the fall quarter, each of approximately 70 Industrial Management students was given five of the above forms, and was told that the checked specimens represented all of the nonaverage performances that he, "the superior," had observed on the part of the five subordinates during the past 6 months, and he was to indicate his overall evaluation of each of the five performances by checking on a scale either, Excellent, Above Average, Average, Below Average, or Poor. Each was asked to pretend that he was the supervisor

of the five men whose separate performances were represented by the five forms. The "observation" score for each form was the number of specimens of effective performance checked or the number of specimens of ineffective performance checked or the difference between these. It was found that increments in evaluation did not grow as the logarithm of the number of specimens checked or the logarithm of the differences. Hence, Fechner's law did not obtain.

About this time an article appeared by S. S. Stevens (1960) in which the number of continua to which the power function applied, as regards the psychophysical relation, was shown to have increased from 14 as reported in a previous article to better than 20. This constituted the stimulus for investigation of the power function as the proper expression for the relation between performance observation and evaluation.

When the mean rating for each of the possible specimen scores in the above study was computed and plotted against the specimen scores on log-log paper, a very good straight line fit was obtained—indicating that equal ratios in the number of performance specimens observed resulted in equal ratios in performance evaluations. In Figure 1 it is seen that this is true both for specimens of effective and ineffective performance. In this figure and in all that follow, each point plotted is based on a minimum of nine observations. The corresponding abscissa value is the midpoint of the interval which contained at least nine observations. Thus the "psychophysical law" was found to obtain for both the case of effective specimens and the case of ineffective specimens when either occurred alone. It then appeared

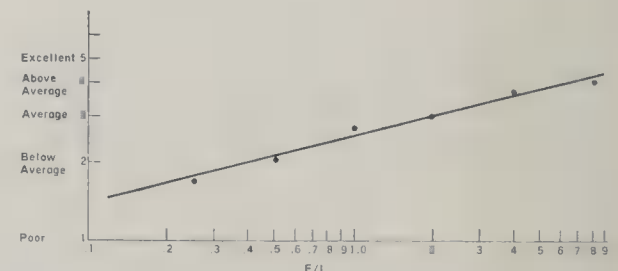
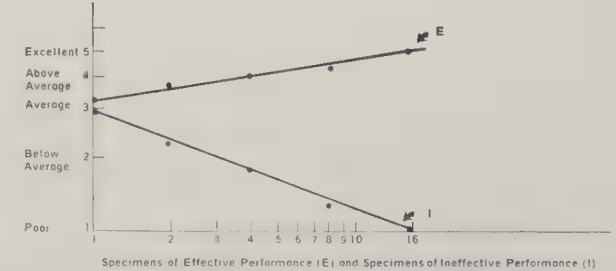


FIG. 1. Effective and ineffective specimens versus evaluation (simulated managerial performance evaluation).

FIG. 2. Ratio of effective to ineffective specimens (E/I) versus evaluation (simulation data).

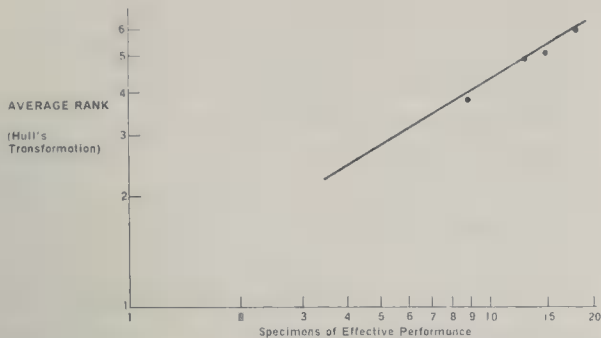


FIG. 3. Effective specimens versus rankings (apprentice linemen).

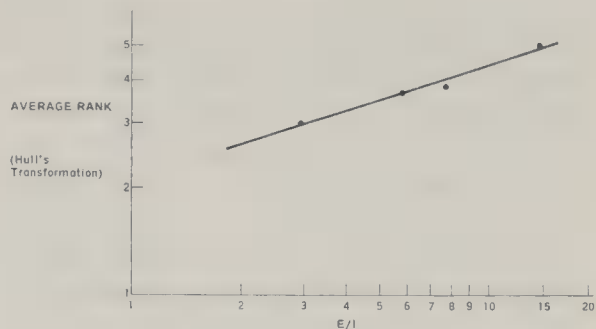


FIG. 4. Ratio of effective to ineffective specimens (E/I) versus evaluation (apprentice linemen).

that the reasonable way to combine specimens of effective and ineffective performance in those cases where both occurred, would be the difference between their logarithms rather than the logarithm of their difference. But this yielded the logarithm of their quotient, so for every case where both effective and ineffective specimens occurred together, the quotient obtained from dividing the number of specimens of effective performance by the number of specimens of ineffective performance was plotted against the resulting evaluation on log-log paper. This also yielded a good fit to a straight line as is seen in Figure 2. It appeared, therefore, that in the case of these experimental data, the psychophysical law quite clearly obtained.

It was possible to extend the scope of this study by reanalyzing some data collected in the past, and by collecting additional data.

In 1951, the writer, along with some colleagues at the University of Tennessee and Personnel Officers from TVA, interviewed every supervisor of apprentice linemen in the Valley, first for the purpose of collecting performance specimens and later for the purpose of obtaining an evaluation of the performance of each of approximately 50 apprentice linemen involved in the study. On this latter trip, each supervisor was handed one performance specimen at a time and was asked to indicate to which if any of the apprentices who had worked under him that specimen pertained. It was possible, thereby, to determine for each apprentice his total number of specimens of effective and ineffective performance. Following this the supervisor was asked to *rank* all the apprentices who had served under him. The

plot on log-log paper of the observations versus evaluations (specimen score versus rank) is given in Figures 3 and 4. In this study as well as the studies which follow, in which actual ratings are obtained, the frequency of ratings where *only* specimens of ineffective performance are reported is too small to permit analysis. Hence, it was not possible to plot these values. It is seen that a reasonably good fit obtains when the average observation value is computed for each man separately and this is plotted against his average rank assignment. It should be noted that the observations and evaluations are not experimentally independent in the studies reported in which actual ratings are used since the raters knew the individuals whose performance specimens they are reporting. However, the effect of this dependence would not appear to be serious since nearly identical results are obtained in the studies involving simulated performance.

In 1957, this investigator collected sets of observations and associated evaluations on approximately 300 executives in a chemical plant in east Tennessee of which the data on 206 were available for the present study. Forms containing 70 specimens of effective managerial performance and 70 specimens of ineffective managerial performance were presented to superiors who checked the specimens they had observed on the part of their subordinates for the previous 12 months and then evaluated their overall performance on a five-point scale as follows: Unusually Effective, Very Effective, Satisfactory, Should Improve, and Must Improve. Figures 5 and 6 present the plot on log-log paper of the observations versus the evaluations. It is seen

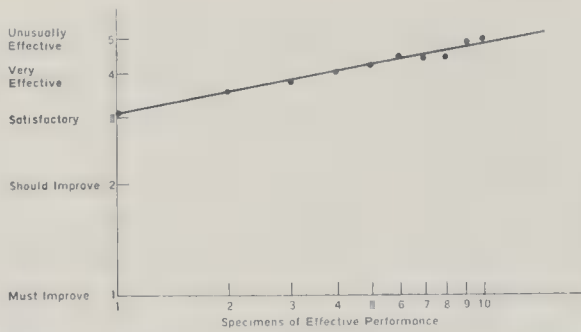


FIG. 5. Effective specimens versus evaluations (chemical plant executives).

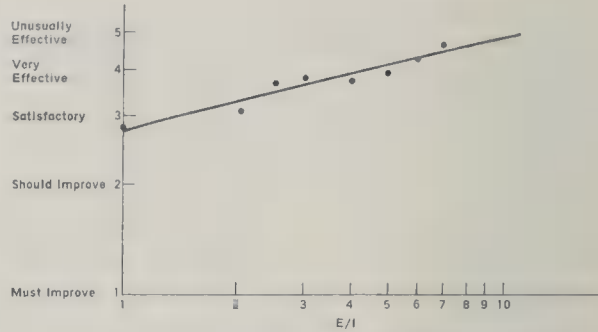


FIG. 6. Ratio of effective to ineffective specimens (E/I) versus evaluation (chemical plant executives).

that the straight line fit is rather good, again indicating conformity to the psychophysical law.

The next set of data was collected over a 5-year period, beginning with an attempt by J. M. Porter, Jr. of the University of Tennessee and this investigator to collect a representative set of performance specimens for the area of supervisory performance called Human Relations Performance. These specimens were collected by questionnaires from a sample of second-line supervisors from manufacturing establishments over the country. These specimens were later edited and a jumbled list of 26 specimens of effective and 36 specimens of ineffective performance was presented in 1960 to supervisors in four plants in east Tennessee. The supervisors were instructed to check the specimens they had observed during the past 6 months on approximately 300 subordinates, and then they made an overall evaluation of the human relations performance on the following scale: Excellent, Above Average, Average, Below Average, and Poor. Figures 7 and 8 present the plot on log-log paper of

the observations versus evaluations. In spite of extreme variability of individual ratings, the mean evaluations are clearly related to the observation values by the psychophysical law.

The next set of data consisted of college students' responses to a set of specimens of teacher classroom behavior. Each student checked the specimens he had observed in class during the past quarter, giving a total of over 800 ratings. Each student was instructed also to evaluate his professor's overall classroom performance on an A, B, C, D, and F scale. Figures 9 and 10 demonstrate that again the psychophysical law described the relationship between observation of performance and evaluation of performance.

At this point it seemed fairly safe to conclude that the judged quality of performance grew as a power function of the number of effective performance specimens observed or as a function of the ratio of the number of specimens of effective to ineffective performance for the performance areas investigated.

However, verification of the theory of

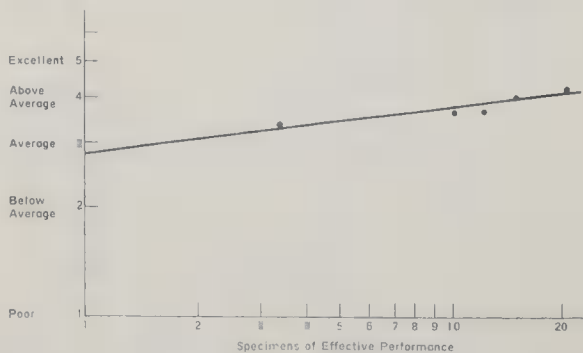


FIG. 7. Effective specimens versus evaluation (supervisory human relations performance).

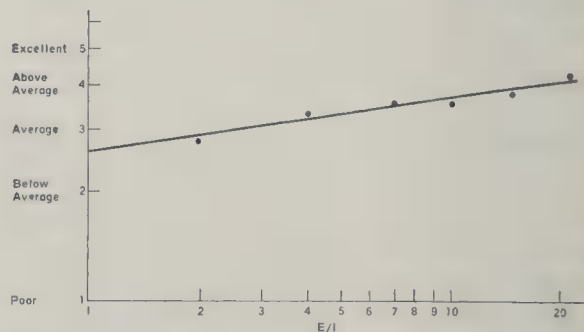


FIG. 8. Ratio of effective to ineffective specimens (E/I) versus evaluation (supervisory human relations performance).

performance evaluation here presented required a further experiment. It had been demonstrated that if a suitable sample of performance specimens was presented to a population of raters, then mean evaluation values could be predicted from mean observation values. The evaluation process did seem to consist basically of the observation of nonaverage effective and/or ineffective performances and responding to the observations in accordance with the psychophysical law. However, there remained the possibility that the sample of performance specimens presented to the raters somehow affected the evaluations, and that if the raters had not been exposed to such a sample, their evaluations would have been other than those obtained when the sample of specimens was presented *prior* to the evaluation. Accordingly, the following experiment was designed to test the effects of presenting the sample of specimens to the rater prior to his evaluation of performance.

Each of a control group consisting of 65 students was presented the 40-item sample of teacher classroom performance specimens referred to above from which they selected and checked the specimens they had observed on the part of two of their professors during the quarter. They were instructed to check the specimens they had observed on the part of each professor and then grade the professor's overall classroom performance as A, B, C, D, or F.

Another 65 students, the experimental group, did likewise except that they evaluated the professors' performance *first* on the same

A to F scale and *then* were presented the sample of specimens for checking. The results of the two rating conditions are given in Figures 11 and 12. It is observed that: (a) The psychophysical law describes all relations between observation and evaluation of performance, i.e., both control and experimental groups for both effective specimens only as well as E/I. (b) Evaluations made prior to exposure to a sample of performance specimens obey the same law as regards their relation to observation as do evaluations made after exposure to a sample of performance specimens.

Two further experiments were conducted for the purpose of more firmly establishing the analogy between performance evaluations and psychophysical processes. In the first, 70 Industrial Management students at the University of Tennessee were each presented five "evaluation" forms and were told to assume that they were executives in a manufacturing plant whose job it was to evaluate the performance of the five supervisors represented by the forms. The observation period was stated as 6 months. On each form was listed a number which was identified as, "Number of observations of very effective performance" and another number which was identified as, "Number of observations of very ineffective performance." At the bottom of the form was stated, "Your over-all evaluation of this supervisor's performance during the rating period is:" the student then checked one of the following: Excellent, Above Average, Below Average, or Poor. Thus 350 ratings were obtained

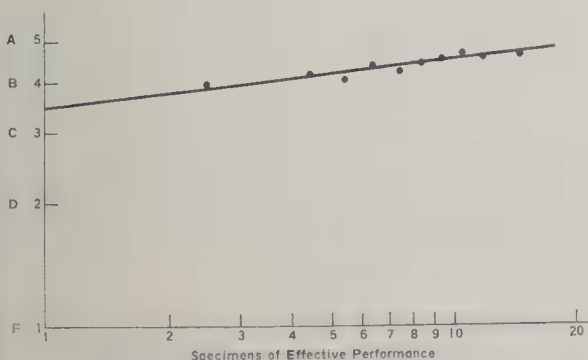


FIG. 9. Number of specimens of effective performance versus evaluation (college teacher classroom performance).

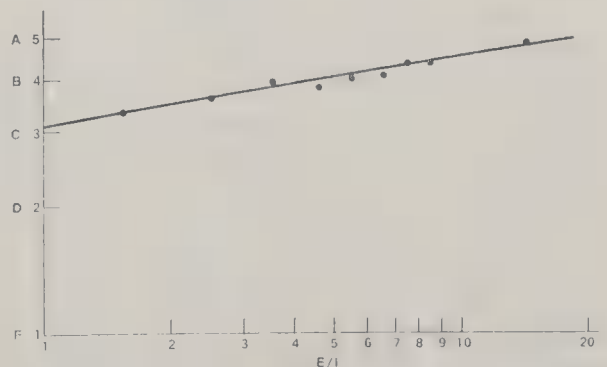


FIG. 10. Ratio of effective to ineffective specimens (E/I) versus evaluation (college teacher classroom performance).

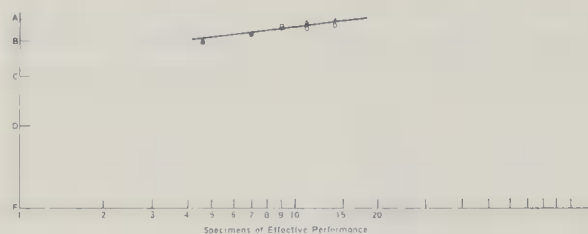


FIG. 11. Specimens of effective performance versus evaluation for control group (Δ) and experimental group (o).

under the condition of having the student observe only the *number* of specimens of effective and ineffective performance rather than observing the actual specimens. Hence the students were presented absolutely no information about the performance except the number of specimens of effective performance and the number of specimens of ineffective performance. By this means, a physical scale was accomplished for the observation values. The psychological magnitude scale (evaluations) was accomplished by means of the above five-point rating scale.

Figure 13 demonstrates that under these conditions the psychophysical law continues to describe the relationship between the physical and psychological magnitudes.

The next experiment was conducted for the purpose of determining whether or not the performance evaluation process had an additional characteristic common to prothetic continua as described by Stevens (1960). The study was conducted by David Gray (1962). The performance of nine hypothetical employees was represented by nine different combinations of specimens of effective and ineffective performance. The range was from 6 Es and 1 I to 2 Es and 4 Is. This yielded an E/I ratio range of .5 to 6.0. Forty supervisors in a local chemical plant were divided into 2 groups of 20 each. Each supervisor was presented nine separate "forms" on which were listed in jumbled order the specimens of effective and ineffective performance. One group evaluated the nine "performances" using a category scale. They were given the following instructions:

The performance specimens on this form represent a very good employee while the specimens on this

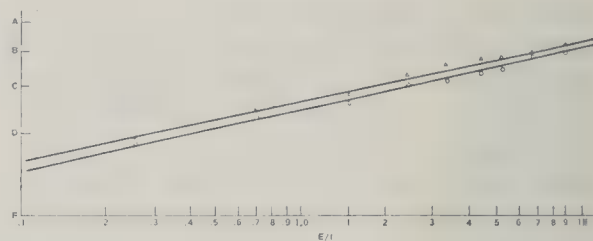


FIG. 12. Ratio of effective to ineffective performances (E/I) versus evaluation for control group (Δ) and experimental group (o).

form represent a poor employee. Nine hypothetical employees are represented by observed performance incidents listed on the other nine forms. The forms are identified as Jack, Bill, etc. Using desirability as a subordinate employee as the criterion, rate these forms into categories so that the interval between them appears to be equal.

While they were not instructed to avoid ties, every supervisor placed the nine "employees" in nine separate categories.

The other group evaluated the same nine performances by means of a constant-sum method as described by Torgerson (1958). In this method, each set of specimens is compared to every other set by dividing 100 points between each of the 9 (9-1)/2 or 36 separate pairs. The mean point values are computed for each set of specimens and converted to logarithms. For each pair the difference between the logarithms is computed, and a matrix is set up which permits the averaging of those differences for each set of specimens. The antilog of this average is the desired scale equivalent and is a ratio scale. In the present study, the following directions were given each of the 20 supervisors in the second group:

A hypothetical employee is represented by each of these nine forms which consist of observed performance specimens. Your rating of these employees is to be based on desirability as a subordinate employee. Compare two employees and divide 100 points between them so as to reflect your evaluation of the employees. Compare each employee to every other employee and enter your judgments in the score sheet in the appropriate box. Just as an example, Bill may be rated at 60 and Dan 40.

The scale values for the category rating were obtained simply by assigning values one through nine to the nine categories used and then computing the mean value for each set of specimens.

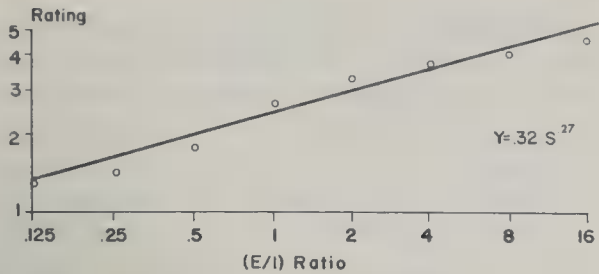


FIG. 13. Log-log plot of ratings versus (E/I) ratios for the numbers presentation.

Thus for each of nine performances (sets of specimens) there was obtained both a category scale value and a ratio scale value. According to Stevens a near universal attribute of prothetic continua is that when category scale values on such continua are plotted against ratio scale values, there results a curve concave downward. Figure 14 presents the plot of the category and ratio scale values for the nine sets of specimens. (Only seven points are plotted since in two cases different combinations of effective and ineffective specimens produced the same E/I ratio. The plotted value is the mean of the scale values for the two sets.) The characteristic concave downward curve is ob-

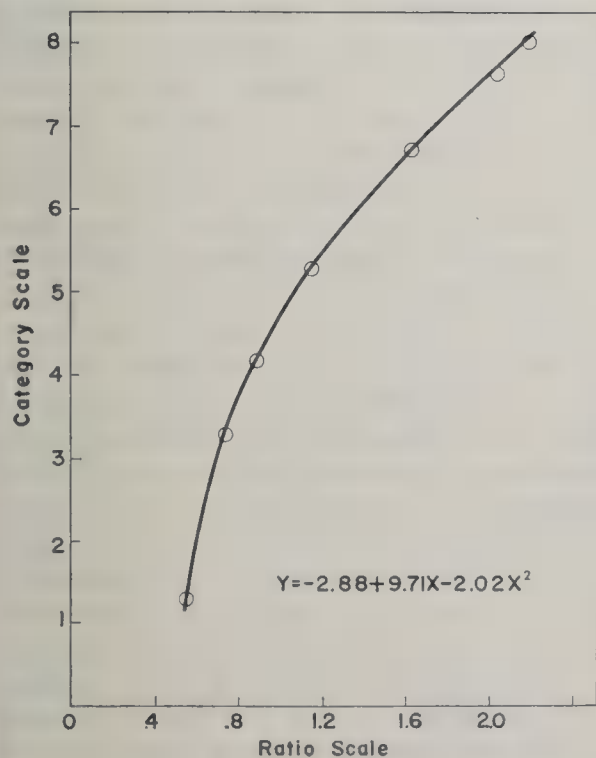


FIG. 14. Category versus ratio scales.

served thus satisfying that criterion for a prothetic continuum. Figure 15 presents the ratio scale values and the E/I values plotted on log-log coordinates. While one reversal occurred, the linear trend is apparent, indicating that the psychophysical law describes the relationship between the E/I values and the ratio scale values.

The results of the present studies seem to offer a new and promising approach to the study of the evaluation process. Having established the nature of the law governing the process, measures of variation from predicted relations are now possible. (Table 1 presents the equations for the data presented above.) It is no longer necessary to rely for our criteria of the "goodness" of a rating method solely on the shape of the distribution or correlations with other criteria of questionable relevance. The variance of estimate for prediction of evaluations of performance from observations provides the needed measure of the precision of the rating process—or if one prefers, the amount of "noise" in the system.¹ The variability of evaluations from predicted values can be studied as a function of a host of factors such as length of observation period, nature of rating categories, effect of spacing of effective and ineffective specimens, and many others. Equally important is the possibility of estimating through these methods the amount of error variance in ratings due to the inconsistency of the performance being evaluated.

¹ Correlations between ratings and E/I for individual raters were found to be: .62 (Beard, 1960), .72 (Hedge, 1960), and .71 (Whatley, 1961).

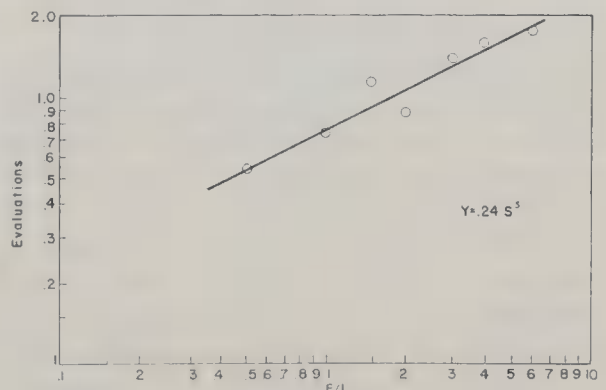


FIG. 15. (E/I) versus evaluations, log-log coordinates.

TABLE 1
EQUATIONS FOR ESTIMATING PERFORMANCE EVALUATION (y) FROM OBSERVATIONS (S)

Area of performance	Equation
Simulated ratings	
S = effective performance only	$y = 3.20S^{.15}$
S = ineffective performance only	$y = 2.90S^{.39}$
S = effective/ineffective	$y = .64S^{.20}$
Apprentice linemen	
S = Effective performance only	$y = 1.08S^{.56}$
S = effective/ineffective	$y = 2.24S^{.29}$
Chemical plant executives	
S = effective performance only	$y = 3.08S^{.19}$
S = effective/ineffective	$y = 2.76S^{.24}$
Supervisory human relations	
S = effective performance only	$y = 2.90S^{.11}$
S = effective/ineffective	$y = 2.64S^{.13}$
Teacher classroom performance ^a	
S = effective performance only	$y = 3.52S^{.12}$
S = effective/ineffective	$y = 2.05S^{.16}$
Teacher classroom performance ^b	
S = effective performance only	$y = 2.48S^{.13}$
S = effective/ineffective	$y = 1.65S^{.20}$
Teacher classroom performance ^c	
S = effective performance only	$y = 2.92S^{.09}$
S = effective/ineffective	$y = 1.40S^{.24}$

^a Classroom ratings made anonymously by students, but turned in to professor.
^b Ratings made outside of class, not seen by the professors rated; exposure to specimens preceding evaluation (see text).
^c Ratings made outside of class; evaluation preceding exposure to specimens.

Among the practical implications of these findings is the possibility of setting confidence limits for evaluations corresponding to the possible observation values. Indeed, it will not even be necessary to ask raters for evaluations once the equations are established. The above studies have been of interest also because of the implications for the study of judgment processes. The kinds of judgments in these studies have typically been considered quite complex as compared to judgments of such things as loudness or intensity of light. Yet one is impressed with their similarity when both are considered in the light of Stevens' model for prothetic processes. Studies are underway at the University of Tennessee in the area of morale,

job evaluation, and attitude measurement to determine whether judgments in these areas behave in a similar fashion. Preliminary findings in the area of morale indicate that self-estimate of morale level is a power function of the ratio obtained by dividing the number of high morale specimens by the number of low morale specimens during a previous 6-month period. (A morale specimen is any occurrence or event which causes a definite shift in level of morale.) In the area of job evaluation average community rates for nine hourly paid key jobs were found to be a power function of judged job value where the latter judgments were obtained by the constant-sum method described earlier.

SUGGESTIONS FOR FURTHER RESEARCH IN
PERFORMANCE EVALUATION

Theoretically there exists for every area of human performance a set of performance specimens such that for every representative sample from the set, the evaluation is determined. Hence, anything less than a perfect correlation between appropriate performance observation measures and measures of evaluation means that some attribute of the observer or the performance observed or their interaction is significantly influencing the transition from observation to evaluation of the performance. In the case of perfect correlation, the error of estimate in predictions from observation to evaluation would be zero. Hence, the magnitude of the error of estimate is a direct measure of the extent to which other than "pure" performance variables are influencing the relationship between observation and evaluation. (It has been established that customary correlation procedures are appropriate if both the observation values and the evaluation values are converted to logarithms.)

As indicated earlier, observer variables might include such factors as experience, recency, idiosyncratic weighting of specimens, length of observation period, relation of rater to ratee, perceived purpose of the evaluation, rater competency, and the like. It remains to discover the attributes of various classes of specimens and the manner of their operation. Three attributes and tentative hy-

potheses concerning the manner of their operation are suggested:

1. Specimen criticality—Performance specimens vary in ascribed weight. Of two samples of observations containing an equal number of specimens of effective performance, the set containing specimens with highest mean weight will result in higher evaluation. The set having the greatest variance among the criticality values will produce the greater error of estimate in predicting evaluations from observations.

2. Specimen factorial dimensionality—The judgment of total performance quality is the weighted sum of the observation values on a set of performance dimensions. The observation of a set of specimens of effective performance (at a given level of criticality) from different performance dimensions will result in a higher evaluation than that resulting from observation of a set all of which load high on the same dimension. Or, if the specimens in both sets are on the same dimension, the difference in evaluation will be a function of the mean factor loadings within each set. The set containing the lower mean factor loading will receive the higher evaluation and the higher error of estimate of evaluations from observations. In the case of sets of observations whose elements differ in factorial dimensionality and specimen criticality, the level of evaluation is a function of the sum of the quotients obtained by dividing each specimen criticality weight by its factor loading.

3. Factorial homogeneity—A given performance specimen is unlikely ever to be factorially pure. However, its factorial *structure* can be represented by the vector obtained by plotting its loadings on the unrotated factors (Cureton, 1960). The similarity of its factorial structure to that of another specimen is measured by the cosine of the angle between the item vectors. It is hypothesized

that a given number of specimens of effective performance (at the same criticality level) which are determined to be factorially homogeneous will result in a higher evaluation and lower error of estimate than a numerically equal set with low factorial homogeneity and whose highest loadings are on the same (rotated) factor.

The above three specimen attributes (criticality, factorial dimensionality, and factorial homogeneity) appear to this writer as the ones most crucial in their effects on the level of evaluations and on the precision with which can be predicted evaluations from observed performance specimens. The hypotheses concerning their manner of operation are presently no more than hypotheses. However, they do represent a few of the many which can be stated precisely and tested unequivocally. Hence, the direction of such research seems promising as a means of furthering an understanding of the evaluation process.

REFERENCES

- BEARD, B. An analysis of the relationship between the number of effective and ineffective performance incidents and the rated quality of performance. Unpublished master's thesis, University of Tennessee, 1960.
- CURETON, E. E. Dimensions of airman morale. *USAF WADD tech. Note*, 1960, No. 60-137.
- GRAY, D. The methods of psychophysics applied to the evaluation of employees. Unpublished master's thesis, University of Tennessee, 1962.
- HEDGE, W. D. Functions mediating the transition from observation of performance to evaluation of performance: A preliminary study. Unpublished master's thesis, University of Tennessee, 1960.
- STEVENS, S. S. The psychophysics of sensory function. *Amer. Scientist*, 1960, **48**, 226-253.
- TORGERSON, W. S. *Theory and method of scaling*. New York: Wiley, 1958.
- WHATLEY, H. G. The effect of the manner of presentation of stimuli upon the relationship between the observation of performance and the evaluation of performance. Unpublished master's thesis, University of Tennessee, 1961.

(Received January 17, 1962)

PERSONALITY FACTORS IN THE SELECTION OF CIVILIANS FOR ISOLATED NORTHERN STATIONS¹

MORGAN W. WRIGHT, GEORGE C. SISLER, AND JOANNE CHYLINSKI

University of Manitoba

An investigation of personality characteristics associated with favorable adjustment to northern isolated living, and the usefulness of psychological tests in the selection of personnel for northern posting. 197 electronic technicians already screened on other psychological tests completed MMPI, Edwards Personal and Brainard Preference tests, and General Information and Arctic Interest questionnaires prior to 1 year of isolation duty on the Mid-Canada Line. Adequacy of work and social adjustment was associated with 11 of 35 test variables and 3 of 26 questionnaire items. The discriminating function of the MMPI was reduced by the use of the K correction. It was suggested that despite the highly select nature of the sample used, the test battery has potential value in the selection of civilians to work in the far north.

The rapid development of the Canadian far north in the past two decades has brought into prominence the question of the capability of individuals to withstand the personal and social isolation which may be associated with Arctic living. It is generally agreed that working in isolated northern outposts involves types of stress and hardship not common in urban communities. Hence, as much selection as is possible should be exercised to protect morale by excluding those who are unable to tolerate working for long periods under these conditions.

A number of studies using military personnel as subjects have been undertaken by the American Armed Forces to determine those personality characteristics which distinguish the soldier or airman who adjusts successfully to the far north (Debons, 1950; Eilbert, Glaser, & Hanes, 1957; McCollum, 1950, 1951). Broadly speaking, their findings indicate that the best predictor of adjustment to an Arctic military environment is an individual's previous history of adjustment to his job and social environment. However, as Willis (1960) has pointed out, very little systematic research has been conducted on adaptation to northern living using a non-military group.

The present study was designed to investi-

gate those personality characteristics which contribute toward favorable adjustment to isolated living. A further aim was to determine the usefulness of certain psychological tests in the selection of civilian personnel for northern posting.

METHOD

Subjects

The first group studied was comprised of Bell Telephone Company of Canada electronic technicians who had volunteered for a 1-year tour of isolation duty on the Mid-Canada Line. The company itself conducted an initial screening procedure which included a detailed interview, personal history questionnaire, medical examination, aptitude tests (Iowa Physics Aptitude Test, Circuit Tracing Test), personality test (Gordon Personal Profile), and a test of general learning ability (Wonderlic Personnel Test), on the basis of which they selected 292 applicants from a total of 547, for a 6-12 month preparatory training course. Of the 292 candidates participating in the training course, which included briefing regarding northern isolated living as well as technical instruction, the company made a final selection of 197 for posting. Their initial screening and training procedures were thus designed to provide an appraisal of the candidate's general stability and physical fitness, his aptitude for technical work, and his attitude towards the anticipated isolation. The successful candidates were consequently a highly select and technically competent group.

Materials and Procedure

Immediately prior to their departures to the northern field sites the 197 employees completed, in

¹ The research reported in this paper was supported by Grant No. 9425-07 from the Defence Research Board, Department of National Defence, Ottawa, Canada.

groups of 40-50, a battery of five tests and questionnaires. The psychological tests employed were the MMPI (Hathaway & McKinley, 1951), Edwards Personal Preference Schedule (Edwards, 1959), and Brainard Occupational Preference Inventory (Brainard & Brainard, 1956). In choosing these questionnaires the investigators were guided by the past work of Fulkerson, Freud, and Raynor (1958) and Sells (1955, 1956) in selecting Air Force personnel for duty under stress. The three standardized tests were supplemented by the following:

General Information Form. This was comprised of 12 items relating to the background and history of the subject (age, education, marital status, number of years married, children, marriage plans if single, number of previous positions, type of home residence, organizational memberships, interests and hobbies, types of vacations preferred, and number of previous illnesses).

Arctic Interest Form. This was composed of 14 items relating to attitude toward and interest in the north (Do you look forward to working in the north? Do you know what kind of work you are going to do in the north? Is the north an area which you have always wanted to visit? Have you ever been in the north before? How long would you like to remain in the north? How long do you feel a person should remain in the north? Do you believe griping can be reduced by helping people out or leaving them alone?), items relating to sports and traveling (Do you like to hunt and fish? Do you like to travel? Do you like winter sports? Do you like to get off by yourself? Do you or would you like extended camping trips with small groups of men?), as well as items relating to previous community size (size of the community in which subject grew up, size of the community in which subject lived prior to going north).

A tour of isolation duty refers to small community living for a period of 1 year on the Mid-Canada Line located along the fifty-fifth parallel. This defense line is organized into six sections each of which is maintained by approximately 150 men. A section is comprised of one "main base" where the majority of the men live, and 10 or 12 outlying "Doppler sites" which are operated by a minimum of two, but occasionally four to eight men. Conditions of isolation were not the same for each individual in terms of Doppler experience, however, they all spent the same overall length of time in the north.

Following the completion of a 1-year tour of duty, each individual's adjustment to isolated living was appraised by the area superintendent under whose supervision he had worked. The definition of "adjustment" employed in this study is similar to that used by Eilbert, Glaser, and Hanes (1957) who define Arctic adjustment as "adequacy of overall job performance and ability to get along with other people and co-workers" (p. 2). In keeping with this definition, an individual rating form was modeled after one employed by the above authors, providing for the scoring of individuals on each of two

parameters: work and social adjustment. The subject was rated for work adequacy as: unsatisfactory, poor, average, good, or excellent. His general social adjustment was rated as: unsatisfactory, adequate, or good. The supervisors were instructed to assess each individual impartially, with the understanding that the ratings would be treated confidentially and would have no influence on the employee's or the rater's future with the company. A total of 170 ratings were received, 27 subjects not being rated because of inability to contact the two supervisors under whom they had worked. It is not believed that this omission introduced a systematic bias in the data collected.

ANALYSES AND RESULTS

A "top adjustment" group of 64 subjects and a "bottom adjustment" group of 52 subjects were selected from the total of 170 on whom ratings were received. The top group was comprised of those rated excellent or good on Parameter 1 plus good on Parameter 2. The bottom group was comprised of those rated unsatisfactory, poor, or average on Parameter 1, plus unsatisfactory or adequate on Parameter 2. This procedure excluded those subjects whose scores on the two parameters were not consistent, i.e., subjects attaining a score of either excellent or good on Parameter 1 plus a score of unsatisfactory or adequate on Parameter 2, or subjects attaining a score of unsatisfactory, poor, or average on Parameter 1 plus a score of good on Parameter 2.

The *t* and *F* tests of significance were first employed to assess the differences between the means and between the variances of the top and bottom adjustment groups on the 35 variables of the three psychological tests. A one-tailed test of significance was used to assess the variables of the MMPI only, since predictions regarding the direction of their differences could be made on the basis of past research (Debons, 1950; Fulkerson et al., 1958; McCollum, 1950, 1951). A two-tailed test was otherwise employed. Using not-*K*-corrected scores of the MMPI, seven variables from this test, three variables from the Edwards, and one variable from the Brainard show differences at better than the .05 level of confidence. These 11 variables have been brought together in Table 1. It is apparent from Column 5 of Table 1 that of these 11 variables with significant differences

TABLE 1
COMPARISON OF TOP AND BOTTOM ADJUSTMENT GROUPS ON THE VARIABLES OF THE MMPI, EDWARDS
PERSONAL PREFERENCE SCHEDULE, AND BRAINARD OCCUPATIONAL PREFERENCE INVENTORY

Variable	Top adjustment group		Bottom adjustment group		Variance ratio	Mean difference	CR	Point-biserial <i>r</i>
	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>				
MMPI								
<i>K</i> score	15.140	3.572	13.826	4.427	1.536*	1.314	1.731*	.163
<i>Es</i>	52.290	3.433	50.538	4.853	1.998**	1.752	2.250*	.208
not <i>K</i> corrected								
<i>Hs</i>	2.062	2.107	2.846	2.645	1.575*	−0.784	−1.777*	−.164
<i>Pd</i>	12.906	3.593	14.846	4.070	1.283	−1.940	−2.724**	−.246
<i>Pt</i>	6.312	4.635	8.923	6.492	1.962**	−2.611	−2.523**	−.229
<i>Sc</i>	6.078	4.270	9.077	6.180	2.095**	−2.999	−3.082**	−.278
<i>Ma</i>	14.812	4.440	16.327	4.954	1.245	−1.515	−1.735*	−.160
<i>K</i> corrected								
<i>Hs</i> & .5 <i>K</i>	9.687	2.905	9.923	2.542	1.305	−0.236	−0.466	
<i>Pd</i> & .4 <i>K</i>	18.625	4.173	20.211	4.547	1.187	−1.586	−1.937*	
<i>Pt</i> & 1 <i>K</i>	21.440	4.767	22.723	6.114	1.645*	−1.283	−1.238	
<i>Sc</i> & 1 <i>K</i>	21.297	3.828	22.904	5.549	2.101**	−1.607	−1.840*	
<i>Ma</i> & .2 <i>K</i>	17.765	4.464	19.250	4.773	1.143	−1.485	−1.716*	
Edwards								
Deference	14.296	3.079	12.730	4.375	2.018**	1.566	2.180*	.207
Orderliness	14.734	4.372	12.326	5.711	1.706*	2.408	2.502*	.233
Aggression	11.171	3.869	13.673	5.462	1.992**	−2.502	−2.783**	−.260
Brainard								
Professional	57.540	11.533	62.540	11.183	1.064	−5.000	−2.311*	−.214

Note.—Only variables which significantly discriminate top-bottom adjustment criterion are shown.
* $p \leq .05$.
** $p \leq .01$.

between the means, 8 also show significant differences in variances at better than the .05 level. The 11 variables were subjected to point biserial correlation with the top-bottom dichotomous variable. The values obtained are given in Column 8. Also shown in Table 1 are *K*-corrected values of MMPI scores. With the addition of the *K* correction, only five variables from the MMPI show differences at better than the .05 level of confidence.

The 14 items of the Arctic Interest Form were dichotomized as closely as possible to the median score of the total group, and in eight instances where approximation of the median was not possible, trichotomization was employed to yield three equal divisions. Chi square values were obtained for the top and bottom adjustment groups. Two items showed significant differences at better than the .02 level of confidence, size of the community where the subject grew up ($\chi^2 = 12.892$, $df = 2$, $p = < .01$), and size of the

community in which the subject has lived most of his life ($\chi^2 = 8.727$, $df = 2$, $p = < .02$).

Two items from the General Information Form, those relating to the age and education of the subject, were treated with the *t* test of significance. The difference between the mean ages of the top and bottom adjustment groups was found to be significant at better than the .05 level of confidence ($t = 2.458$, $df = 114$). The mean age of the top group was 23.790 ($SD = 4.859$) compared with a mean age of 21.812 ($SD = 3.098$) for the bottom group. The remainder of the variables of this form were treated with the chi square test. No significant relationships were found.

DISCUSSION

Personality Factors and Adjustment

MMPI scores of the bottom adjustment group are significantly higher than those of the top adjustment group on the Hypochon-

driasis, Psychopathic Deviate, Psychasthenic, Schizoid, and Manic scales; and lower on the Suppressor (*K*) and Ego Strength scales. This latter scale, developed by Barron (1953), measures personal resourcefulness and capacity for adaptability. It is interesting to note the degree of consistency in these findings with earlier studies, none of which employed the *K* correction. Debons (1950) found that on the MMPI scales administered prior to an Alaskan tour of duty, a group of infantrymen who, following a period in the Arctic, rated themselves as "less able" to endure a further tour in the north, attained significantly higher scores than a "more able" group on the Hypochondriasis, Depression, Hysteric, Psychopathic Deviate, Schizoid, and Validity scales. McCollum (1950), in a study of a group of 100 airmen rated as a "low morale" group, found that the highest frequency of scores falling above a criterion of two standard deviations above the mean occurred on the Psychopathic Deviate and Manic scales, followed by a high frequency of deviations on the Schizoid, Psychasthenic, and Depression scales. He also found (McCollum, 1951) that a group of military offenders attained significantly higher scores than a group of nonoffenders on the Psychopathic Deviate, Manic, Paranoid, Hypochondriasis, Hysteric, and Validity scales. Fulkerson et al. (1958) found that the scores of a poorly adjusted group of pilots were higher than those of a well adjusted group of pilots on the Validity, Hypochondriasis, Depression, Hysteric, Psychopathic Deviate, and Psychasthenic scales, and lower on the *K* scale. In all the above mentioned studies as well as the present one, the Psychopathic Deviate scale differentiated the two groups. In four of the five studies the Hypochondriasis, and in three of the five studies the Depression, Hysteric, Schizoid, Manic, and Validity scales were associated with less favorable adjustment.

The findings of the present study indicate that those variables associated with anti-social behavior and psychotic tendencies are related to less efficient functioning in conditions of isolation. They suggest that individuals whose personalities are characterized by mood instability, restlessness, and over-

activity; individuals who are inclined toward oversensitivity, seclusiveness, and emotional shallowness; individuals whose anxieties are self-directed, as well as individuals exhibiting limited resourcefulness and capacity for dealing with stressful situations as measured by the Ego Strength scale, will experience some difficulty in making a favorable adjustment to northern isolated living.

In their study, Fulkerson et al. (1958) indicated that the use of the *K* correction had the effect of reducing the discriminability of the MMPI scales. This same observation was made in the present study. Since the mean *K* score for the top adjustment group was significantly higher than the mean *K* score for the bottom group, the addition of the *K* correction elevated the scores of the top adjustment group most, thus making the differences between the mean scores of the two groups smaller. It is apparent from Table 1 that the addition of this factor reduces the significance of the Hypochondriasis and Psychasthenic scales to below the .05 level of confidence. This observation raises the question of the value of the use of the *K* correction in studies such as the present one, since the direction which the mean *K* score takes appears to affect the discriminating function of the entire scale. Heilbrun (1961) and others (King & Schiller, 1959; Smith, 1959) have suggested that within a normal population the *K* score is related to general psychological adjustment rather than being a measure of defensiveness in test taking attitude (Meehl & Hathaway, 1946). Heilbrun compared the *K* scores of maladjusted male and female college counseling service clients with a sample of male and female normal college students, but was able to demonstrate a significant difference in *K* scores only between his normal females and a subgroup of more seriously maladjusted female counseling service clients, the normal female group having the higher mean *K* score. His results only partially supported his contention that the *K* scale of the MMPI is a measure of psychological adjustment in a normal population. The finding in the present study that the mean *K* score of the well adjusted group of male technicians was significantly higher than that of the poorly adjusted group is

consistent with Heilbrun's hypothesis. Further, the observation that the addition of the *K* correction to MMPI scales reduces the predictive usefulness of the test supports his statement (Heilbrun, 1961) that "if *K* is positively correlated with psychological adjustment for normal subjects and is not a measure of defensiveness, the *K* correction would appear to be operating in direct opposition to test validity" (p. 486).

Three variables of the Edwards test were observed to differentiate the better functioning from the poorer functioning groups. The top adjustment group attained higher scores on the Deference and Orderliness variables and lower scores on the Aggression variable. Thus, the poorly adjusted individual feels less need for organization and orderliness, is less accepting of leadership, and shows a more critical and aggressive attitude.

One variable of the Brainard test was found to differentiate the two groups. This variable is related to interest in professional occupations, the well adjusted group indicating less preference for professional activity.

Two variables of the Arctic Interest Form which are related to and deal with the urban-rural characteristics of the subject's developmental background were both significant in their relationship to top and bottom adjustment. A greater proportion of the better functioning group were of rural backgrounds, but for some years prior to their postings lived in large urban areas. This finding is not inconsistent with that of Eilbert et al. (1957) who found that members of their low adjustment group grew up in large urban communities. Stunkel, Tye, and Yaukey (1952) reported that individuals who stated they preferred city activities were more likely to be in the group judged to be poorly adjusted. Although these findings are not conclusive they suggest that a rural background better prepares a person to withstand the rigors of northern living than a purely urban one.

The only item of the General Information Form reliably discriminating the two groups was that relating to age. It would appear that the probability of making an unsatisfactory adjustment to isolated northern living is greater the younger the individual.

Predictive Value of the Test Battery

The question arises whether the battery of three psychological tests employed in this study could be effectively used for the selection of personnel to work in the north. The degree of relationship between the 11 significant variables and the adjustment criterion shown in Column 8 of Table 1 would suggest that the variables may be of limited predictive usefulness. The potential efficiency of the battery in screening out poor employment risks can be assessed by evaluating scores of all subjects against an arbitrarily selected criterion of maladjustment using the 11 significant variables. The criterion selected by the authors was a score of two standard deviations or greater from the total group mean ($N = 116$) in the direction of poor adjustment on at least one of the 11 significant variables. Using this criterion, 50 subjects, from the total of the 170 initially tested and rated, attained scores indicating maladjustment and would therefore have been predicted to make an unfavorable adjustment to northern isolated living. Of these 50 subjects, 24 (48%) were subsequently rated in the bottom adjustment category, 7 (14%) were rated in the top adjustment category, and the remaining 19 (38%) were in the inconsistently rated or middle group. This would mean that to have screened out 48% of the individuals who reacted unfavorably to northern isolated living, 14% who made a satisfactory adjustment would also have been disqualified as well as 38% who made neither a good nor a poor adjustment. In terms of the total group of 170 subjects, 46% of the bottom adjustment group (24 of 52 subjects), 35% of the middle group (19 of 54 subjects), and 11% of the top adjustment group (7 of 64 subjects) would have been excluded. This same criterion would also have excluded three of the seven individuals who failed to complete their year's tour of duty because of difficulty in adapting to isolated living conditions. Thus in terms of the present study there is evidence that the percentage of civilians making a good work and social adjustment to the north can be increased by using this test battery in addition to other selection procedures. It should be pointed out that the

group used in this study was a highly select one because of the Bell Telephone Company's preselection procedures that were already discussed. The predictive value of this test battery would therefore be applicable only to other similarly select groups. To explore this problem further a second larger and more heterogeneous sample is currently being investigated.

It may also be concluded that there is relatively little evidence in this research to suggest that biographical information can discriminate between men who adjust well to an isolated setting and those who adjust poorly. As well, special interests which may be thought to be related to northern adjustment, or predispositions toward liking the north, appear to have little predictive significance. However, the finding that the two factors, the age of an individual and the urban or rural characteristics of his development background are associated with successful adaptation, warrants further consideration.

REFERENCES

- BARRON, F. An ego-strength scale which predicts response to psychotherapy. *J. consult. Psychol.*, 1953, 17, 327-333.
- BRAINARD, P. P., & BRAINARD, R. I. *Brainard Occupational Preference Inventory manual, Form R*. New York: Psychological Corporation, 1956.
- DEBONS, A. Survey of human adjustment problems in the northern latitudes: Personality predispositions of infantrymen as related to their motivation to endure tour in Alaska. *USAF Arctic Aeromed. Lab.*, 1950, Proj. No. 21-01-022, Part 3-E.
- EDWARDS, A. L. *Edwards Personal Preference Schedule manual*. (Rev. ed.) New York: Psychological Corporation, 1959.
- EILBERT, L. R., GLASER, R., & HANES, R. M. Research on the feasibility of selection of personnel for duty at isolated stations. *USAF Personnel Train. Res. Cent. tech. Rep.*, 1957, No. 57-4.
- FULKERSON, S. C., FREUD, S. L., & RAYNOR, G. H. The use of the MMPI in psychological evaluation of pilots. *J. aviat. Med.*, 1958, 29, 122-129.
- HATHAWAY, S. R., & MCKINLEY, M. D. *Minnesota Multiphasic Personality Inventory*. (Rev. ed.) New York: Psychological Corporation, 1951.
- HEILBRUN, A. B., JR. The psychological significance of the MMPI K scale in a normal population. *J. consult. Psychol.*, 1961, 25, 486-491.
- KING, G. F., & SCHILLER, M. A research note on the K scale of the MMPI and "defensiveness." *J. clin. Psychol.*, 1959, 15, 305-306.
- MCCOLLUM, E. L. Survey of human adjustment problems in the northern latitudes: Relationships between low morale and personality structures. *USAF Arctic Aeromed. Lab.*, 1950, Proj. No. 21-01-022, Part 1-D.
- MCCOLLUM, E. L. Survey of human adjustment problems in the northern latitudes: A study of military offenders. *USAF Arctic Aeromed. Lab.*, 1951, Proj. No. 21-01-022, Part 1-F.
- MEEHL, P. E., & HATHAWAY, S. R. The K factor as a suppressor variable in the MMPI. *J. appl. Psychol.*, 1946, 30, 525-564.
- SELLS, S. B. Development of a personality test battery for psychiatric screening of flying personnel. *J. aviat. Med.*, 1955, 26, 35-45.
- SELLS, S. B. Further developments in adaptability screening of flying personnel. *J. aviat. Med.*, 1956, 27, 440-451.
- SMITH, E. E. Defensiveness, insight, and the K scale. *J. consult. Psychol.*, 1959, 23, 275-277.
- STUNKEL, EVA R., TYE, V. M., & YAUKEY, D. W. Validation of experimental selection instruments for Arctic service. *USA TAGO Personnel Res. Sect. Rep.*, 1949, No. 335.
- WILLIS, J. S. Mental health in the north. *Med. serv. J. Canada*, 1960, 16(8), 689-720.

(Received January 29, 1962)

THE EFFECT OF GROUP PARTICIPATION ON BRAINSTORMING EFFECTIVENESS FOR TWO INDUSTRIAL SAMPLES¹

MARVIN D. DUNNETTE, JOHN CAMPBELL, AND KAY JAASTAD²

University of Minnesota

Problems were presented for brainstorming to 48 research scientists and 48 advertising personnel employed with the Minnesota Mining and Manufacturing Co. Within a counterbalanced experimental design, each S brainstormed certain problems individually and other equated problems as a member of a 4-man team. Individuals produced not only more ideas than groups, but they accomplished this without sacrificing quality. The net superiority of individual performance over group participation is highlighted by the fact that 23 of 24 groups produced a larger number of different ideas under the individual condition. The superiority of individual brainstorming over group brainstorming was relatively greater when it was preceded by group participation. Apparently, group participation is accompanied by certain inhibitory influences even under conditions (e.g., brainstorming) which place a moratorium on all criticism.

Proponents of brainstorming, as an idea eliciting or problem solving technique, emphasize the value of group participation as a facilitating factor in producing ideas. For example, Osborn (1957) concludes on the basis of experiments conducted at the University of Buffalo, that "the average person can think up twice as many ideas when working with a group than when working alone" pp. 228-229). He adds that a combination of group and individual effort is probably best, but he fails to specify the exact nature of the optimal combination. A widely cited study by Taylor, Berry, and Block (1957) suggested that group participation actually inhibits the potential ideational output of individuals. Taylor et al. presented three different problems to 96 Yale juniors and seniors who had previously worked together in small group discussion sections. Forty-eight of the subjects were divided into 12 real groups of 4 men each; the other 48 brainstormed the problems alone. The number of

different ideas produced by the real groups was compared with the number produced by so-called "nominal" groups formed after the experiment by randomly dividing the 48 individual subjects into 12 groups of 4 each. Scores of nominal groups represented, therefore, the expected level of achievement if actual group participation neither inhibits nor facilitates creativeness during brainstorming. For each of the three problems presented, the nominal groups produced an average of nearly twice as many different ideas as the real group. Taylor et al. concluded, therefore, that group participation actually has an inhibiting influence on creative thinking during brainstorming—a conclusion radically different from the one by Osborn cited earlier.

Our purpose has been to repeat the Taylor et al. study among two different occupational groups: research scientists and advertising men. Further, we have modified the design of the Taylor study in order to allow subjects to participate in *both* individual and group brainstorming sessions. We believe our results lend support to the conclusions reached by Taylor and his associates and that they also help to define the conditions for the optimal combination of group and individual effort mentioned by Osborn.

¹ Research reported here was supported by grant from the Graduate School Research Fund of the University of Minnesota.

Appreciation is extended to Raymond O. Collier of the University of Minnesota who helped greatly in planning the statistical analyses of the experimental data.

² Now employed with McKinsey and Company, Chicago, Illinois.

METHOD

The subjects of the experiment were 48 research personnel from one of the larger laboratories of Minnesota Mining and Manufacturing Co. (3M) and 48 persons employed with 3M's central staff Advertising Department. Our choice of research and advertising personnel was based on our hypothesis that advertising personnel would more likely profit from any facilitating influence of group interaction and that research persons would more likely be inhibited by group participation; thus, we hypothesized opposite effects of group interaction in the two groups. Each of the two sets of subjects was divided into 12 groups of 4 men each. The assignment was not random. Instead, persons were placed together who had worked together and who were well acquainted with one another. In no case, however, were persons of differing job levels placed in the same group; among the researchers no one with supervisory responsibilities participated in the study. Among the advertisers, different supervisory levels were represented but never in the same group, thereby reducing any possible inhibiting effects due to differing status levels within a group. In addition, no persons with advanced degrees participated in the study; the researchers included only persons with BA or equivalent degrees. The range of education represented in the advertising group was from high school through college.

Each subject in Taylor's study participated in only one of the two experimental conditions. Such an experimental design does not allow a test of possible effects of prior group experience on individual brainstorming behavior or vice versa, and it also, of course, depends on randomization to effect an equating of any individual differences in brainstorming ability between subjects in the two experimental conditions. Because of these factors, our experiment allowed each subject to take part in both group and individual brainstorming sessions. The experimental design shown in Table 1 was used separately for the researchers and the advertisers. It will be noted that individual and group performance, the two problem sets, and order of participation are counterbalanced.

The problems used in our study were the same as those used by Taylor et al. with the addition of a fourth, the People problem. The four problems are stated below:

Thumbs problem. We do not think this is likely to happen, but imagine for a moment what would happen if everyone after 1960 had an extra thumb on each hand. This extra thumb will be built just as the present one is, but located on the other side of the hand. It faces inward, so that it can press against the fingers, just as the regular thumb does now. Here is the question: What practical benefits or difficulties will arise when people start having this extra thumb?

Education problem. Because of the rapidly increasing birthrate beginning in the 1940s, it is now clear that by 1970 public school enrollment will be

TABLE 1

DESIGN OF THE EXPERIMENT

Group	Individuals	Order	
		First	Second
A	1, 2, 3, 4	I,2	G,2
B	5, 6, 7, 8		
C	9, 10, 11, 12		
D	13, 14, 15, 16	G,1	I,2
E	17, 18, 19, 20		
F	21, 22, 23, 24		
G	25, 26, 27, 28	I,2	G,1
H	29, 30, 31, 32		
I	33, 34, 35, 36		
J	37, 38, 39, 40	G,2	I,1
K	41, 42, 43, 44		
L	45, 46, 47, 48		

Note.—1 = Problem Set 1, Thumbs and Education; 2 = Problem Set 2, People and Tourists; G = Group; I = individual.

very much greater than it is today. In fact, it has been estimated that if the student-teacher ratio were to be maintained at what it is today, 50% of all individuals graduating from college would have to be induced to enter teaching. What different steps might be taken to insure that schools will continue to provide instruction at least equal in effectiveness to that now provided?

People problem. Suppose that discoveries in physiology and nutrition have so affected the diet of American children over a period of 20 years that the average height of Americans at age 20 has increased to 80 inches and the average weight has about doubled. Comparative studies of the growth of children during the last 5 years indicate that the phenomenal change in stature is stabilized so that further increase is not expected. What would be the consequences? What adjustments would this situation require?

Tourists problem. Each year a great many American tourists go to visit Europe. But now suppose that our country wished to get many more European tourists to come to visit America during their vacations. What steps can you suggest that would get more European tourists to come to this country?

The problems were pretested along with several others among University of Minnesota engineering and business administration students in order to select those problems which elicited a large volume and diversity of responses and to equate approximately the idea eliciting qualities of the two problem sets. The pretest problems were presented in counterbalanced order to each subject, and he was allowed 10 minutes for each problem to write his responses. The mean number of responses given by the pretest subjects was not significantly dif-

ferent for Thumbs and People (10.70 and 10.08) or for Education and Tourists (9.62 and 9.80); Thumbs and Education were, therefore, used for Problem Set 1 and People and Tourists for Problem Set 2.

About a week prior to the experiment, the senior author met with the subjects “over coffee” and discussed creative thinking with particular emphasis on brainstorming. The nature and purpose of the forthcoming experiment were explained and questions concerning scheduling and procedural details were answered. The importance of applying the “principles”³ of brainstorming was heavily emphasized, and participants were urged to refrain from discussing their experimental sessions (particularly the problems used) with any of their co-workers who may not yet have participated in the experiment.

The same graduate student (Kay Jaastad) served as experimenter for *all* subjects. She began each experimental session by reading aloud the instructions which emphasized the importance of research study and which restated the techniques and principles of brainstorming. She then presented each problem by first reading it aloud and then distributing dittoed copies to each of the subjects. She allowed time for questions, if any, and then instructed the subjects to “begin brainstorming.” Responses were recorded on a DeJur-Grundig *Stenorette* with conference microphone. Subjects were allowed to spend 15 minutes on each of the problems; in every session, nearly all ideas and solutions had been expressed at the end of 10–12 minutes. The time limit did not in any instance result in cutting off a flow of ideas. On the other hand, it did serve as a stimulus to the rapid and free wheeling expression of ideas and solutions. Each subject participated in both experimental conditions on the same afternoon.⁴ The individual brain-

³ The following principles have been suggested by Osborn and were emphasized by Taylor in his study: the more ideas the better, the wilder the ideas the better, improve or combine ideas already suggested, and do not be critical.

⁴ For example, members of Groups A, B, and C participated first as individuals using Problem Set 1 followed immediately by group brainstorming using Problem Set 2. Those in Groups D, E, and F participated first in the group situation using Prob-

storming condition was carried out by placing subjects in four widely separated offices where each would be free to brainstorm without interruption.⁵

RESULTS

After completing the experimental sessions, the responses of the 96 subjects in both individual and group situations were transcribed. Each idea or solution was placed on one side of a 3 × 5 card, and appropriate identifying information (e.g., problem set, condition, order) was entered on the opposite side. Using the cards, it was an easy (albeit voluminous and time consuming) task to sort responses, delete duplications, rate the quality of ideas, etc. The first step in the analysis of results was simply to compare the number of different ideas or solutions produced by group participation with the number of different ideas or solutions produced by the same group members during the individual brainstorming condition. It should be emphasized that the “score” (number of ideas) under the individual condition includes only *different* ideas. Thus, if two or more members of a group, during their individual sessions, suggested the same idea or solution to a problem, it was counted as only a single contribution to the total score of the nominal group. Comparisons were made between individual and group brain-

lem Set 1 followed immediately by individual brainstorming using Problem Set 2—and so on.

⁵ Unfortunately the individual sessions were not *entirely* free from interruption. In at least one instance, a man answered a phone in the middle of his session. At other times, men were interrupted briefly by passers-by who looked in to ask, “What the heck are you doing?” or to make similar comments. Of course, all such slip-ups worked to the disadvantage of individuals’ achievements in comparison with groups.

TABLE 2
MEAN TOTAL NUMBER OF DIFFERENT IDEAS AND/OR SOLUTIONS TO PROBLEMS BY SUBJECTS
UNDER CONDITIONS OF INDIVIDUAL AND GROUP BRAINSTORMING

Problem	Research personnel		Advertising personnel	
	Individual	Group	Individual	Group
Thumbs and People	78.3	60.9	82.9	59.8
Education and Tourists	62.2	49.3	58.5	37.3
Total	140.5	110.2	141.4	97.1

TABLE 3
ANALYSIS OF VARIANCE: TOTAL NUMBER OF DIFFERENT IDEAS AND/OR SOLUTIONS
TO PROBLEMS BY RESEARCH PERSONNEL UNDER CONDITIONS OF INDIVIDUAL
AND GROUP BRAINSTORMING

Source	Total (both problems of each set)			Thumbs and People problems		Education and Tourists problems	
	<i>df</i>	<i>MS</i>	<i>F</i>	<i>MS</i>	<i>F</i>	<i>MS</i>	<i>F</i>
Between individuals							
Order	1	1001	6.37*	585	9.39*	56	1.57
S × C	1	1291	8.21*	277	4.44	372	10.54*
S × O × C	1	30	0.19	3	0.05	14	0.40
Error (b)	8	157.1		62.3		35.3	
Within individuals							
Condition	1	1380	23.71**	455	15.12**	250	31.64**
C × O	1	40	0.69	10	0.33	10	1.26
Set	1	96	1.65	32	1.05	19	2.40
S × O	1	135	2.33	53	1.74	19	2.40
Error (w)	8	58.3		30.1		7.9	

Note.—S = Set, C = Condition, O = Order, (b) = between, (w) = within.
* *p* < .05.
** *p* < .01.

storming for the total of both problems in each set, for the Thumbs-People problems of each set, and for the Education-Tourists problems of each set. Results are shown in Tables 2, 3, and 4. For all comparisons, the condition effect is highly significant, the indi-

vidual condition yielding a markedly greater number of ideas than group participation. Of the 24 groups, only one failed to produce more ideas under the individual condition than under the group condition, and the difference in this one instance was only 163–162

TABLE 4
ANALYSIS OF VARIANCE: TOTAL NUMBER OF DIFFERENT IDEAS AND/OR SOLUTIONS TO PROBLEMS
BY ADVERTISING PERSONNEL UNDER CONDITIONS OF INDIVIDUAL
AND GROUP BRAINSTORMING

Source	Total (both problems of each set)			Thumbs and People problems		Education and Tourists problems	
	<i>df</i>	<i>MS</i>	<i>F</i>	<i>MS</i>	<i>F</i>	<i>MS</i>	<i>F</i>
Between individuals							
Order	1	274	0.96	153	1.63	18	0.31
S × C	1	293	1.03	78	0.83	69	1.17
S × O × C	1	73	0.29	2	0.02	82	1.39
Error (b)	8	285.1		93.6		58.8	
Within individuals							
Condition	1	2949	93.61**	800	27.78**	677	34.02**
C × O	1	170	5.40*	101	3.51	9	0.45
Set	1	316	10.03*	196	6.81*	14	0.70
S × O	1	125	3.97	24	0.83	39	1.96
Error (w)	8	31.5		28.8		19.9	

Note.—See note in Table 3.
* *p* < .05.
** *p* < .01.

in favor of group participation. Only 5 of the 48 research subjects failed to produce more ideas when working individually than when participating in a group. Clearly, individual brainstorming achieves more ideas than group brainstorming. Our hypothesis that group interaction would facilitate the output of advertising personnel and inhibit the output of research personnel failed to be sustained. Apparently, the inhibiting influence of group participation cuts across the kinds of personal and occupational differences investigated in this study.

It will be noted from Tables 3 and 4 that certain other effects also showed significance. The significant Order effect shown by researchers was most pronounced on the Thumbs and People problems. A larger number of ideas or solutions was produced when subjects experienced the individual brainstorming *after* having experienced the group session than when they “went in cold” to the individual session. Although the Order effect was not significant for the advertisers, the Order \times Condition interaction was. A plot of Individual versus Group scores for each of the two orders shows that the relative superiority of individual brainstorming was greatest when it was preceded by the group session. For example, the mean number of ideas for Individual and Group sessions was 143 and 109, respectively, under Order A (Individual session followed by Group session); the corresponding means under Order B (Individual session preceded by Group session) were 140 and 95. These results help to specify the conditions for combining group and individual effort. Apparently, a brainstorming session (either group or individual) can serve the important function of a “warm up” for subsequent brainstorming. In this study,

the net superiority of individual brainstorming over group brainstorming seems to have been enhanced by such a warm-up session.

Two other statistically significant effects are shown in Tables 3 and 4. The Set effect is significant for the advertising group. Problem Set 1 (Thumbs and Education) gave a higher yield than Problem Set 2 (People and Tourists). The major source of the difference was between the Thumbs and People problems, probably reflecting the fact that they were somewhat less perfectly matched on the basis of the pretests than were the Education and Tourists problems. For the researchers, the interaction effect Set \times Condition is significant. A plot of Individual versus Group scores for each of the two problem sets shows that the superiority of individual brainstorming was greatest (151–91) for Problem Set 1 and that it actually was negligible for Problem Set 2 (130–129). We are at a loss to explain this outcome, particularly in view of the consistent results obtained for *both* problem sets in the advertising group.

Although the total number of ideas and solutions was greater for individual than for group participation, it could be argued that this may have been accompanied by a corresponding decrease in the quality of ideas produced. Two of the scales used by Taylor et al. for rating the quality of ideas were employed in this study. The first two authors used Taylor’s Effectiveness scale⁶ to rate

⁶ Effectiveness scale: 0 = no conceivable contribution to solution of problem. Suggestion impossible of attainment; 1 = very little, if any, contribution to solution of problem; 2 = probably some contribution to solution of problem; 3 = definite minor contribution to solution of problem; and 4 = clearly a major contribution to solution of problem.

TABLE 5
MEAN TOTAL QUALITY SCORES OBTAINED BY SUBJECTS UNDER CONDITIONS
OF INDIVIDUAL AND GROUP BRAINSTORMING

Problem	Research personnel		Advertising personnel	
	Individual	Group	Individual	Group
Thumbs and People	171	131	192	131
Education and Tourists	128	94	116	65
Total	299	225	308	196

TABLE 6
ANALYSIS OF VARIANCE: TOTAL QUALITY SCORES FOR IDEAS AND/OR SOLUTIONS TO PROBLEMS
BY RESEARCH PERSONNEL UNDER CONDITIONS OF INDIVIDUAL AND GROUP BRAINSTORMING

Source	Total (both problems of each set)			Thumbs and People problems		Education and Tourists problems	
	df	MS	F	MS	F	MS	F
Between individuals							
Order	1	6225	8.18*	3385	9.45*	429	3.92
S × C	1	4916	6.46*	1734	4.84	811	7.41*
S × O × C	1	45	.06	0	.00	50	.46
Error (b)	8	761		358		109.5	
Within individuals							
Condition	1	8049	20.17**	2361	11.86**	1691	20.37**
C × O	1	340	.85	103	.52	69	.83
Set	1	995	2.49	425	2.14	119	1.43
S × O	1	172	.43	12	.06	94	1.13
Error (w)	8	399		199		83	

Note.—See note in Table 3.
* $p < .05$.
** $p < .01$.

solutions to the Education and Tourists problems and his Probability scale ⁷ to rate responses to the Thumbs and People problems. Although not *all* responses were rated

⁷ Probability scale: 0 = very highly improbable or clearly impossible; 1 = conceivable, but improbable; 2 = possible; 3 = probable; and 4 = highly probable.

by both investigators, the interrater reliabilities were estimated on the basis of sampling randomly the ideas from each group of subjects for each of the four problems. The resulting coefficients ranged between .54 and .77 with a median value of .66. These reliabilities are *not* impressively high; even so,

TABLE 7
ANALYSIS OF VARIANCE: TOTAL QUALITY SCORES FOR IDEAS AND/OR SOLUTIONS TO PROBLEMS
BY ADVERTISING PERSONNEL UNDER CONDITIONS OF INDIVIDUAL AND GROUP BRAINSTORMING

Source	Total (both problems of each set)			Thumbs and People problems		Education and Tourists problems	
	df	MS	F	MS	F	MS	F
Between individuals							
Order	1	2063	1.67	1162	2.56	133	.63
S × C	1	2959	2.40	1162	2.56	404	1.92
S × O × C	1	1528	1.24	92	.20	883	4.20
Error (b)	8	1235		453		210	
Within individuals							
Condition	1	18676	54.44*	5460	16.25**	3966	60.00**
C × O	1	595	1.73	504	1.50	4	.06
Set	1	115	.34	140	.42	1	.02
S × O	1	10	.03	8	.02	0	.00
Error (w)	8	343		336		66	

Note.—See note in Table 3.
* $p < .05$.
** $p < .01$.

TABLE 8
MEAN QUALITY RATINGS FOR IDEAS AND/OR SOLUTIONS PRODUCED BY SUBJECTS
UNDER CONDITIONS OF INDIVIDUAL AND GROUP BRAINSTORMING

Problem	Research personnel			Advertising personnel		
	Individual	Group	Signifi- cance level ^a	Individual	Group	Significance level ^a
Total (both problems of each set)	2.12	2.04	<i>ns</i>	2.18	2.02	<i>p</i> < .05
Thumbs and People	2.18	2.15	<i>ns</i>	2.32	2.19	<i>ns</i>
Education and Tourists	2.06	1.91	<i>ns</i>	1.96	1.74	<i>p</i> < .01

^a The significance tests were made using the same analysis of variance design shown in Tables 3, 4, 6, and 7. Copies of additional analysis of variance tables may be obtained upon request from the senior author.

the ratings of the second author (which were done for *all* problems in *all* groups) were used to provide a rough index of the quality of each solution proposed. The ratings were summed for all the different ideas produced under each of the conditions. Comparisons of means and statistical tests of significance are shown in Tables 5, 6, and 7. It is apparent that quality was *not* sacrificed for quantity under the condition of individual brainstorming. It is noteworthy that the Set and Condition \times Order effects are not significant among the advertisers for the quality score comparisons. For the advertisers, the Thumbs problem elicited a significantly larger number of responses than the People problem, but the total quality of output did not differ for the two problems.

A remaining question has to do with the mean quality of ideas and solutions produced under the two experimental conditions. Table 8 summarizes the mean quality ratings and the significance levels of the mean differences between individual and group conditions. The values in Table 8 were obtained by dividing the mean total quality score (Table 5) by the mean total number of different ideas (Table 2). It is evident that individuals produce responses of quality equal to or greater than that of the ideas produced in groups. The evidence is clear-cut: brainstorming is most effective when undertaken by individuals working *alone* in an atmosphere free from the apparently inhibiting influences of group interaction.

DISCUSSION

Our results confirm those of Taylor et al. and tend to refute Osborn's argument that

individuals are stimulated by group brainstorming to produce more ideas than when brainstorming alone. Of special interest is our finding that group interaction has an inhibiting influence for advertising people (that area in which brainstorming was developed and where it first came into widespread use) as well as for technical research personnel and for college students at Yale University (Taylor's study). Individuals not only produce *more* ideas when working alone, but they do this without sacrificing quality; indeed our results show that advertising personnel, working as individuals, produced ideas on the Tourists and Education problems of significantly higher mean quality than when they worked in groups. The net superiority of individual performance over group participation for these two sets of industrially employed subjects is highlighted by the fact that 23 of the 24 groups produced a larger number of different ideas under the individual condition. To the extent that we may generalize these findings to future situations, we can state that four persons, attacking a problem individually, and then pooling their efforts will, on the average, produce about 30% more ideas than if they attempted to solve the problem in a group session or meeting.

Our findings also suggest that group participation may be useful in "warming up" for individual brainstorming sessions. Research personnel produced more ideas when individual brainstorming followed group participation than when it preceded it. Advertising men also exhibited relatively greater superiority in the individual sessions when they had been preceded by a group session.

Neither the Taylor study nor our study has identified the exact nature of the inhibiting influence which apparently acts to reduce the productivity of group brainstorming. Taylor et al. suggested, and we concur, on the basis of our observations during these experiments, that a group tends to "fall in a rut" and to pursue the same train of thought. The effect of this is to limit the diversity of approaches to a problem, thereby leading to the production of fewer different ideas. It was also apparent that the output of many individuals who were highly productive when working alone was considerably less in the group situation. In spite of the stimulus of group brainstorming and our specific directive to avoid all criticism, it was apparent that these persons were inhibited simply by the presence of other group members. The cen-

tral idea underlying brainstorming of placing a moratorium on all criticism is a good one. It appears, however, that group participation still contains certain inhibitory influences which are not easily dissipated. The "best bet" for creative thinking in attacking problems seems, therefore, to be the pooled individual efforts of many people with perhaps an initial group session to serve simply as a warm up to their efforts.

REFERENCES

- OSBORN, A. F. *Applied imagination*. (Rev. ed.) New York: Scribner, 1957.
- TAYLOR, D. W., BERRY, P. C., & BLOCK, C. H. Does group participation when using brainstorming facilitate or inhibit creative thinking? Technical Report No. 1, 1957, Yale University, Department of Psychology, Office of Naval Research.

(Received February 2, 1962)

WORK GROUP ATTRIBUTES AND GRIEVANCE ACTIVITY

W. W. RONAN

Mesta Machine Company

5 hypotheses concerning the relation of grievance submission to nature and setting of work activity were tested. The hypotheses were that a particular plant has no effect on grievance submission, number of grievances submitted has no relation to nature of work, number of grievances "won" has no relation to nature of work, grievance goals do not vary by nature of work, rate of grievance submission is not related to nature of work. Results showed grievance submission does vary by plant and certain work groups won more grievances. The other 3 hypotheses were not supported. However, an overall tabulation of grievance activity makes all results somewhat tenuous.

Why do workers submit grievances? This question, if asked in a group of industrial relations people, would probably get different answers equal to the number in the group and, as likely, generate spirited discussion. A recent book by Sayles (1958) has presented information to indicate that a unique answer might be given to the question. In general, Sayles states that work groups take on certain attributes which are related to the nature and setting of the work they perform. On this basis, workers are classified into four groups. They are:

1. Apathetic (A)—those who do dirty, dangerous work involving little or no skill. They may be individual or "crew" jobs where the crew cooperates in performing a common work task. This latter is in contrast to something of the nature of an automobile assembly line. They are also found in "mixed" departments where a product may be machined, assembled, and tested by different persons or groups.

2. Erratic (E)—those whose work involves a major physical component but is easier than that above. Primarily are assembly line jobs where work flows to another group but with little direct communication between the groups. The work tends to be controlled more by the workers than that of the A groups.

3. Strategic (S)—semiskilled work involving sufficient judgment so that work standards are rather loose. Primarily work is individual, such as grinder operators. They are higher status jobs and of relatively more importance in the organization.

4. Conservative (C)—highly skilled and usually, but not always, crafts. Involve criti-

cal and scarce skills where the work is strictly individual and requires long training or experience. In general, individual and noninterdependent work.

As described in the book, one of the measurable attributes of such work groups is the amount and nature of their grievance activity. E and A—unpredictable grievance rates; C—low grievance rates; S and C—much better "thought out"; i.e., legalistic grievances; and the latter groups also tend to press for new concessions rather than attempting to preserve customary rights through opposition.

If the technical nature of work does directly influence the behavior of work groups, it is an important consideration in establishing jobs, selecting and training supervisors, plant layout, operations management, and industrial relations generally. This study is an attempt to evaluate Sayles' observations through an analysis of formal grievance activity in one company.

PROCEDURE

All the formal written grievances from the two company plants for the years, July 1957 through June 1961, were obtained and analyzed. The grievances thus cover a period of 4 years. One plant in the organization was put into operation in 1957 whereas the other plant has been in operation for over 50 years.

The older plant has approximately three times as many "shop employees" ($N = 2,850$, as described for this study; as the newer plant, $N = 850$) and the latter is unionized, while the former is not. However, in both a formal grievance procedure is prescribed, in the one plant through union representatives, in the other, as prescribed in the company "policy book."

The company manufactures heavy processing equipment for the metals industry. As such, it is a "job shop" since virtually every piece of equipment is unique. This consideration makes it likely that the E group is much underrepresented in this study since such groups are typically found in interdependent work groups or short assembly lines. There is some work of this nature in the company but it is a relatively minor part. For this reason the established employee categories are not completely satisfactory but, for the purposes of this study, were used. These two groups are of the A-E and S-C types as previously described. For the purposes of this study the latter group is composed of all the skilled crafts, cranemen, inspectors, and such other skilled workmen. The former group is composed of all other persons holding relatively less skilled jobs. In terms of size, the two groups are almost exactly one-half the total number of production employees—on the average about 1,850 in each group.

RESULTS

Hypotheses

From the above background, the following hypotheses were derived and subjected to test:

- 1. The particular plant has no effect on the number of grievances submitted.
- 2. There is no difference in number of grievances submitted by the two groups.
- 3. There is no difference between the two groups in number of grievances "won."
- 4. There is no difference in the goals of the grievances submitted by the two groups.
- 5. There is no difference between the two groups in rate of grievance submission.

To evaluate the first hypothesis, the grievances from each of the two plants plus the theoretical number, on the basis of number of employees at each, were tabulated.

A chi square test of these data indicated an extreme significance level. The null hypothesis is rejected, i.e., the particular plant does have an effect on the number of grievances submitted.

TABLE 1
GRIEVANCES BY PLANT

Distribution	New plant	Old plant
Actual	63	38
Theoretical	23.20	77.80

TABLE 2
GRIEVANCES BY WORK GROUPS

Distribution	C-S group	A-E group
Actual	50	51
Theoretical	47.1	53.9

The second hypothesis is based upon the information from Sayles' book that a C-S group should, in general, submit fewer grievances than an A-E group. Actual and theoretical numbers of grievances are shown in Table 2.

The null hypothesis is accepted on the basis of a chi square value of .16. The number of grievances submitted does not differ by groups.

Due to the fact that the C-S group plan their grievances better, they should "win" more grievances than the other group. Table 3 shows a tabulation for evaluating this hypothesis.

The results show that the C-S group won-lost record does not differ from chance, while the A-E record does between the 1-2% level. On this basis the null hypothesis is rejected in favor of the finding that a difference in winning or losing grievances does exist between the two groups.

The C-S group is presumed to differ from the other group in the goals of the grievances submitted. Generally speaking, they are seeking concessions or variations in terms of job content, policies, practices, etc., while the other group tends to be more defensive or negativistic in that it is opposing actions or expected events. After categorizing on the basis of grievance content, an evaluation of the fourth hypothesis is shown in Table 4.

TABLE 3
WON-LOST GRIEVANCES BY WORK GROUPS

Distribution	C-S group		A-E group	
	Won	Lost	Won	Lost
Actual	23	27	14	37
Theoretical	25	25	25.5	25.5
	$\chi^2 = .15$		$\chi^2 = 5.45$	

TABLE 4
GRIEVANCE CONTENT BY WORK GROUPS

Distribution	C-S group		A-E group	
	New	Oppose	New	Oppose
Actual	23	27	21	30
Theoretical	25	25	25.5	25.5
	$\chi^2 = .15$		$\chi^2 = .83$	

The chi square results show that neither distribution differs significantly from chance. The null hypothesis of no difference in goals of grievances is accepted.

The fifth hypothesis is based upon Sayles' observations that the C-S group tends to keep a rather steady rate of grievance submission in contrast to a more erratic rate of submission by the other group. An evaluation of this hypothesis is taken from the data in Table 5.

The hypothesis was evaluated by computing a *t* value between the standard deviations for the two distributions. The standard deviations were: C-S—1.106 and A-E—3.344. A *t* value of 1.8 was obtained which is not high enough to indicate a significant difference, with consequent acceptance of the null hypothesis.

In summary, evaluations of the above tabulated data show that a particular plant can have a higher grievance rate; the C-S

TABLE 5
YEARLY GRIEVANCE RATE BY WORK GROUPS

Year	Number of grievances	
	C-S group	A-E group
1957-58	15	17
1958-59	13	9
1959-60	12	15
1960-61	10	10
Total	50	51

groups win more grievances; groups show no difference in rate or number of grievances submitted, and no difference in goals of submitted grievances. However, a more detailed tabulation of the data casts some doubt on such overall evaluations of the hypotheses.

Table 6 is a more detailed tabulation of the data used in this study.

The previous finding that more grievances are submitted in a "newer" plant seems to be supported by the data in Table 6, particularly since the old plant has three times as many employees as the new. Ignoring the year 1957-58 still shows the newer plant with a much higher grievance rate per employee.

However, in all the other cases it appears likely that the null hypothesis is the more probable. With regard to number of grievances, in some years one group is higher

TABLE 6
YEARLY GRIEVANCE RATES AT PLANTS BY WORK GROUPS

Plant	Year	C-S grievances		A-E grievances		Total
		Won	Lost	Won	Lost	
Old	1957-58	1	4	0	3	8
	1958-59	1	6	0	5	12
	1959-60	4	3	1	3	11
	1960-61	1	1	2	3	7
	Subtotal	7	14	3	14	38
New	1957-58	9	1	5	9	24
	1958-59	5	1	2	2	10
	1959-60	1	4	2	9	16
	1960-61	1	7	2	3	13
	Subtotal	16	13	11	23	63
Total		23	27	14	37	101

and the next the other. Rate of grievance submission and "winning or losing" show the same general pattern.

CONCLUSION

While the overall results support some of Sayles' observations, the more detailed breakdown makes such inferences somewhat doubtful. It is, however, pertinent to point out that the described relationships may be distorted by the fact of a new plant—the fact that over one-third of the total grievances at that plant were submitted in the first year would seem to indicate that "newness," as such, is an important variable in grievance submission and, possibly, the "normal" grievance pattern has not yet been established.

This study is also limited in both time and scope. Longer time periods or broader observational criteria might show more conclusive results, while including several companies more clearly subtending all four work groups could also change the findings.

Finally, it is possible that Sayles' findings might better be tested by using the results of morale and attitude studies as a criterion. Certainly, the term "grievance" is not defined only by those of a formal nature and broader criteria might well give more conclusive results.

In summary, the broad findings of Sayles tested against a restricted criterion to assess the relation between formal grievances and work group attributes has given tenuous results. However, the amount and detail of documentation in Sayles' book make it seem likely that a criterion of more scope might show entirely different results. In any case, the importance of the implications in the book makes further tests of such implications desirable.

REFERENCE

- SAYLES, L. R. *Behavior of industrial work groups*. New York: Wiley, 1958.

(Received February 12, 1962)

AN EMPIRICALLY-DERIVED MANAGERIAL KEY FOR THE CALIFORNIA PSYCHOLOGICAL INVENTORY

LEONARD D. GOODSTEIN

University of Iowa

AND WILLIAM J. SCHRADER¹

*Personnel Research Branch, Civilian Personnel Field Agency, Ordnance Field Activity,
Rock Island, Illinois*

Chi square comparisons of the responses of 603 managers and supervisors with those of 1748 men-in-general indicated that 206 of the 480 California Psychological Inventory (CPI) items reliably ($p < .01$) differentiated the 2 groups. All of the managerial and men-in-general CPI protocols were then scored using these 206 items as a Managerial key. This key not only reliably differentiated the total managerial group from the men-in-general group but also differentiated personnel at 3 different levels of management: top management, middle management, and first-line supervision (all p 's $< .01$). This CPI Managerial scale also significantly correlated ($r = .233$) with ratings of success within the total management group and within the top and middle management subgroups (r 's = .254 and .267, respectively). These results were compared with results of other recent personnel research and the implications discussed.

Personality characteristics are given a major role in most theories of occupational and professional success, especially at the managerial and supervisory level where such characteristics are frequently regarded as more important than skill or technical knowledge. The implementations of this point of view, insofar as the development of techniques for measuring such personality characteristics are concerned, has not been very successful. A perusal of such sources as Buros (1953, 1959) indicated that the few attempts in the past to develop psychological tests to measure such personality factors have not been successful.

One rather promising psychometric technique for assessing personality characteristics is the California Psychological Inventory (CPI) developed by Gough (1957) for use with relatively normal individuals found in schools, colleges, business, and industry. Designed to evaluate the positive aspects of personality which are important for social living and social interaction and that have wide and pervasive applicability to human behavior, the CPI is a self-administering, 480-

item, true-false response, personality inventory. The purpose of the present study was to determine the usefulness of the CPI in identifying those personality characteristics associated with managerial and supervisory success in a large industrial organization.

While several of the 18 published CPI scales, such as Dominance and Capacity for Status, may be regarded as potentially useful in determining these personality characteristics, a recent study by Mahoney, Jerdee, and Nash (1960) reported that, of 13 unnamed CPI scales, only the Dominance scale was actually related to rated managerial success. Further, it may be noted that none of the 18 published CPI scales was developed especially for the purpose of assessing managerial potential and none had utilized a large industrial sample. Consequently it was decided to develop an empirical Managerial scale or key, based upon a comparison of the actual responses of a large sample of managers and supervisors with those of a large sample of nonmanagerial personnel. It was expected that such a scale, developed in this fashion, might be a useful instrument in the assessment of managerial potential within similar large industrial organizations and also

¹ The authors are indebted to Kenneth B. Hoyt for his assistance in the initial planning of this study and to Nathan Jaspán for his assistance in the statistical analysis of the data.

might be useful in the actual selection of management personnel.

METHOD

Subjects. The CPI was administered to a sample of 603 male civilian supervisors and managers (the management sample) at eight United States Army Ordnance Corps Field Service Depots as part of a large scale personnel research project. These depots are responsible for the storage, supply, maintenance, and repair of various items of military equipment and employ large numbers of skilled, semiskilled, and unskilled civilian workers. The management sample ranged in age from 26 to 68 with a mean age of 42.7 ($SD = 7.8$); the education level of the management sample ranged from 5 to 20 years of formal schooling with a mean educational level of 11.4 ($SD = 2.2$).

The management sample was subdivided into three subsamples: (a) top management ($N = 106$) which included those persons at the Assistant Division Chief level or higher, those with fairly broad management responsibilities; (b) middle management ($N = 245$) which included those persons employed in management and supervision below that of top management and above line supervision; (c) line supervision ($N = 252$) which included those persons immediately directing the activities of line workers. CPI protocols were also available on a random sample of 1,748 ordnance depot men-in-general, a nonprofessional, nonscientific, and non-supervisory sample which had been collected as part of an earlier ordnance personnel research project. The men-in-general sample ranged in age from 20 to 62 with a mean age of 40.9 ($SD = 10.2$) while the educational level ranged from 2 to 20 years of formal schooling with a mean of 9.7 ($SD = 2.5$).

Ratings. Each individual in the management sample was rated by his immediate supervisor on the adequacy of his on-the-job performance. These ratings were secured in individual interviews with trained personnel technicians, using a fairly typical five-step, 20-attribute (e.g., effectiveness in communication, effectiveness in planning, etc.) rating form, which yielded a single, averaged numerical rating. These ratings were then used to validate the final CPI Managerial key.

RESULTS

The total managerial sample was randomly split into two halves with 302 persons in one half and 301 in the other half. (There were no differences in the proportions of the three managerial subgroups in the two samples.) The differences in the relative frequencies of "yes" and "no" responses to each of the 480 CPI items between each of the two managerial half-samples and the men-in-general sample were determined and statistically analyzed

by means of chi square tests. Those items which yielded statistically significant chi square values at the 1% level of confidence (6.64 for 1 *df* in the same direction for *both* managerial half-samples) were then combined into a single Managerial key² which consisted of 206 of the 480 items.

Each of the managerial and men-in-general CPI protocols was then scored by this Managerial key. Each item answered in the keyed direction was arbitrarily credited with one point while each item answered opposite to the key was scored minus one. To avoid negative scores 500 was added to the total score for each individual. The means, *SDs*, and ranges of the Managerial scale scores for the three managerial subsamples, the total managerial sample, and the men-in-general sample are presented in Table 1.

As can be noted from Table 1 the mean CPI Managerial scale scores were substantially higher for the total management sample than for the men-in-general sample, although there was considerable overlap of the two distributions. A *t* test indicated that this difference was highly reliable ($p < .01$). Within the management sample, the Managerial scale scores are highest at the top management level, next highest at the middle management level, and lowest at the line supervision level. An overall *F* test and subsequent *t* tests indicated that all of the differences among these managerial subgroups were highly statistically reliable (all $ps < .01$), although again there was considerable overlap among the three distributions. Thus it would appear that the CPI Managerial scale reliably differentiates managerial personnel from nonmanagerial personnel and, further, differentiates managerial personnel at the three levels of management.

Pearson product-moment correlations were

² The CPI item numbers and the direction of the keyed response for the items included in the Managerial key are presented in a 1-page table which has been deposited with the American Documentation Institute. Order Document No. 7195 from ADI Auxiliary Publications Project, Photoduplication Service, Library of Congress; Washington 25, D. C., remitting in advance \$1.25 for microfilm or \$1.25 for photocopies. Makes checks payable to: Chief, Photoduplication Service, Library of Congress.

TABLE 1
MEANS, *SD*s, AND RANGES OF THE CPI MANAGERIAL SCALE SCORES FOR THE SEVERAL GROUPS, TOGETHER WITH THE CORRELATIONS WITH THE CRITERION RATINGS

Group	CPI Managerial key				Correlation with rating
	<i>N</i>	<i>M</i>	<i>SD</i>	Range	
Top management	106	626.6	27.3	526-678	.254**
Middle management	245	605.4	36.1	496-680	.267**
Line supervision	252	594.2	36.2	486-664	.115
Total management	603	604.5	36.7	486-680	.233**
Men-in-general	1,748	550.9	47.6	394-662	—

** $p < .01$.

then computed between the CPI Managerial scale scores and the independently-obtained criterion ratings of on-the-job performance for the total managerial sample and each of the subsamples. The obtained correlations are also presented in Table 1. The magnitude of the obtained correlations, although small, indicated that the CPI Managerial scale scores were positively and significantly ($p < .01$) correlated with the criterion ratings for the total management group and in the top management and middle management subgroups. In the line supervision subgroup, however, the correlation between the Managerial scale scores and the criterion ratings was not statistically reliable. Thus it would seem that the CPI Managerial scale not only differentiates the managerial subgroups from each other and nonmanagerial personnel but also is a statistically reliable index of rated success in management at all but the first-line supervisory level.

DISCUSSION

The results of the present study tend to support the contention that an empirical approach to the assessment of managerial potential using the CPI is a useful one. While the differences between the total managerial and men-in-general groups may be spuriously large as a result of having done the item validation with the same groups, the highly significant differences among the mean scale scores for the three managerial subgroups as well as the correlations with the performance ratings suggest that this new Managerial scale has some promise for measuring mana-

gerial potential within groups and perhaps also may prove useful for individual prediction. Naturally, the results are limited by the characteristics of the present sample and require cross-validation before any broad generalizations can be made.

Items from all of the 18 published scales are included in this new CPI Managerial scale with more than half the items from the Tolerance (72% of the items), Achievement via Independence (62%), Dominance (52%), Capacity for Status, Self-Acceptance, and Achievement via Conformance (all 50%) scales included. All of the other published scales contributed a smaller percentage of items with the Femininity scale the only one yielding a substantial number of items scored in the reversed direction (69%). Successful managerial personnel would therefore appear to be nonauthoritarian, achievement oriented, dominant, high drive, communicative, self-acceptant, and nonfeminine individuals (Gough, 1957, pp. 23-27). Of course, it must be recognized that these personality test differences between the managers and non-managers are probably truly confounded with a number of variables, including education, intelligence, socioeconomic status, etc.

The obtained results suggest that there may be significant differences in personality characteristics not only between managers and nonmanagers but also among managers at different levels of responsibility. These results, like those reported by Ghiselli (1959), suggest that the usual differentiation of personnel into managers or line workers is not sufficient, and that the differences within the manage-

ment group constitute an often neglected source of variance. Unlike Ghiselli, however, there was a clear differentiation between the first-line supervisory group and the men-in-general group in the present study. The failure of the Managerial scale to reliably predict success *within* the first-line supervisory group does support Ghiselli's suggestion that personality factors may operate differently at the several levels of management. In this case it might be argued that success in first-line supervision is mainly determined by technical skill and knowledge which is relatively independent of personality factors while, at the upper levels of management, such personality-related variables as organizing and directing, planning, decision making,

etc., become important. Additional research on this important management problem seems highly desirable.

REFERENCES

- BUROS, O. K. (Ed.) *The fourth mental measurements yearbook*. Highland Park, N. J.: Gryphon, 1953.
- BUROS, O. K. (Ed.) *The fifth mental measurements yearbook*. Highland Park, N. J.: Gryphon, 1959.
- GHISELLI, E. E. Traits differentiating management personnel. *Personnel Psychol.*, 1959, 12, 535-544.
- GOUGH, H. G. *Manual for the California Psychological Inventory*. Palo Alto, Calif.: Consulting Psychologists Press, 1957.
- MAHONEY, T. A., JERDEE, T. H., & NASH, A. N. Predicting managerial effectiveness. *Personnel Psychol.*, 1960, 13, 147-163.

(Received February 16, 1962)

REANALYSIS OF EXPERIMENTAL HALO EFFECTS

DONALD M. JOHNSON¹

Michigan State University

Data on ratings of individuals, obtained under 2 conditions of judgment and published in 1956, were reanalyzed by a more complete analysis of variance. The usual interaction between raters and individuals, called a halo effect, was found but it was not influenced by judgment conditions intended to maximize it. Hence, the evidence for halo effect due to judging operations remains questionable.

A study of halo effects published several years ago (Johnson & Vidulich, 1956) used Guilford's (1954) adaptation of the analysis of variance to compare ratings obtained under conditions designed to maximize halo and conditions designed to minimize halo. Recently, at the urging of several correspondents, the data have been analyzed more thoroughly and an error has been corrected. Since no other research by this powerful technique has appeared in the interim, the more complete analysis is presented briefly.

Two groups, each composed of 18 college students, rated five individuals on five traits, using a nine-point scale. To maximize halo one group rated one individual each day on all traits, and the other group, to minimize halo, rated all individuals each day on one trait. The details and a limited analysis of variance are available in the original report. The present analysis is concerned with the particular individuals and traits included in the experiment, while the raters are taken as a random sample of college students, and the error terms for the *F* tests were chosen accordingly, following the discussion of Bennett and Franklin (1954). Table 1 shows the complete analysis for the two conditions separately, and Table 2 shows the analysis for the two conditions combined.

Table 1 shows that the main effects are significant under both conditions, as by the previous analysis. The interactions between individuals and raters and between traits and raters are both significant under both conditions. By the previous, less sensitive analysis the $I \times R$ interaction was significant only under the maximization condition and the

$T \times R$ interaction only under the minimization condition. The $I \times T$ interaction, overlooked in the previous analysis, is large and significant under both conditions. Willingham and Jones (1958) have shown that this interaction is inversely related to the older measure of halo, the intertrait correlation. It might also be called a measure of differential discrimination since it measures the agreement by the raters in assigning certain traits to certain individuals. Senator Joseph McCarthy, for example, was rated low on kindness and Sir Winston Churchill was rated high on intelligence.

Comparing the two conditions, it is clear from Table 1 that the rater variances are about the same. The variance due to individuals is larger under the maximization condition, as expected, but the *F* ratio ($69.04/42.00 = 1.64$) is not significant. Going back to the original method of describing halo, the mean correlation among raters, over traits, of individuals is estimated from Stanley's (1961) Equation 31 as .40 for the maximization condition and .42 for the minimization condition. The $I \times R$ interaction, called the relative halo effect, is also not significantly different from one condition to another. Hence these results do not permit the conclusion that judgment under conditions intended to minimize halo actually reduced the halo effect.

The trait variance appears to be larger under the minimization condition, although the *F* ratio of 4.33 is not significant with only 4 and 4 degrees of freedom. However the $T \times C$ interaction of Table 2 is significant at the .01 level when tested against the $T \times R$ term. And the $T \times R$ interaction under the minimization condition when compared with the maximization condition yields an *F* ratio

¹ The statistical advice of Terrence M. Allen is gratefully acknowledged.

TABLE 1

ANALYSIS OF VARIANCE OF RATINGS OF FIVE PROMINENT INDIVIDUALS ON FIVE TRAITS BY TWO GROUPS OF 18 RATERS EACH UNDER CONDITIONS INTENDED TO MAXIMIZE AND MINIMIZE HALO EFFECTS

Source	df	Maximization			Minimization		
		SS	MS	F	SS	MS	F
Raters	17	251.78	14.81	10.50**	243.14	14.30	7.33**
Individuals	4	276.16	69.04	12.83**	167.99	42.00	10.63**
I × R	68	365.68	5.38	3.82**	268.65	3.95	2.03**
Traits	4	217.54	54.39	26.79**	50.19	12.55	2.38*
T × R	68	138.70	2.03	1.44*	358.85	5.28	2.71**
I × T	16	292.66	18.29	12.97**	334.92	20.93	10.73**
I × T × R	272	384.78	1.41		529.84	1.95	
Total	449	1927.30			1953.58		

* $p \leq .05$.
** $p \leq .01$.

of 2.60, significant at the .01 level. Hence it appears that the attempt to minimize halo by having the raters make their judgments on one trait each day actually decreased the trait variance and increased the trait-rater interaction. We can only conclude that manipulation of judgment did have an effect, though not the expected one.

“Halo effect” implies an effect of the rating operation, a tendency for global rather than analytic judgment, but the older evidence from correlation coefficients could be interpreted as due to objective variations between the individuals or the public information about them (Johnson, 1945). The I × R variance appears to be better evidence for

halo, but even this variance could be due to objective variations in the information available to the different judges rather than to the judging operation. Just as the I variance may be an effect of mass communication, the I × R variance may be an effect of selective communication. It was these weaknesses in the evidence for halo as a phenomenon of judgment, pointed out in 1945, that prompted the 1956 attempt to show that experimental manipulation of the rating conditions, with information held constant, would change this interaction variance. Since the expected differences were not significant, we are back where we were in 1945, with no good evidence for halo effects in ratings.

REFERENCES

BENNETT, C. A., & FRANKLIN, N. L. *Statistical analysis in chemistry and the chemical industry*. New York: Wiley, 1954.

GUILFORD, J. P. *Psychometric methods*. (2nd ed.) New York: McGraw-Hill, 1954.

JOHNSON, D. M. A systematic treatment of judgment. *Psychol. Bull.*, 1945, **41**, 193-224.

JOHNSON, D. M., & VIDULICH, R. N. Experimental manipulation of the halo effect. *J. appl. Psychol.*, 1956, **40**, 130-134.

STANLEY, J. C. Analysis of unreplicated three-way classifications, with applications to rater bias and trait independence. *Psychometrika*, 1961, **26**, 205-219.

WILLINGHAM, W. W., & JONES, M. B. On the identification of halo through analysis of variance. *Educ. psychol. Measmt.*, 1958, **18**, 403-407.

(Received February 16, 1962)

TABLE 2

OVERALL ANALYSIS OF VARIANCE OF THE TWO GROUPS OF TABLE 1

Source	df	SS	MS	F
Conditions	1	11.04	11.04	.76
Raters within Conditions	34	494.92	14.56	8.68**
Individuals	4	436.76	109.19	23.43
I × C	4	6.60	1.65	.35
I × R in C	136	634.33	4.66	2.77**
Traits	4	215.26	53.82	14.70**
T × C	4	51.67	12.92	3.53**
T × R in C	136	497.55	3.66	2.18**
I × T	16	584.08	36.51	21.73**
I × T × C	16	43.50	2.72	1.62
I × T × R in C	544	914.62	1.68	
Total	899	3890.33		

** $p \leq .01$.

THE EFFECT OF CORRECT RESPONSE LOCATION ON THE DIFFICULTY LEVEL OF MULTIPLE-CHOICE QUESTIONS ¹

ARTHUR MARCUS ²

University of Massachusetts

The influence of position response sets on a multiple-choice achievement test was investigated. Data on 434 students were obtained from 4 alternate test forms of 100 items each. The arrangement of correct choices and distractors was randomized throughout the test by a scheme which allowed each position an equal number of correct choices. The correct choice for each item appeared in a different position on each of the forms. Results indicate that objective multiple-choice tests are relatively free of position preferences. With this type of test it appears that position response sets are negligible and certainly not a significant source of invalidity. It is suggested that the position of the most plausible distractor more logically accounts for any significant response bias than does a position preference.

Several investigations of position response set, in multiple-choice tests, have yielded contradictory results. Cronbach (1950) concluded that objective multiple-choice tests are relatively free from position response sets. He analyzed test papers from both the Henmon-Nelson Test of Mental Ability and the Ohio State Psychological Examination and found no evidence for the response set. However, McNamara and Weitzman (1945) report that a tendency does exist among test takers to favor certain response positions to the neglect of others. Their findings suggest that the difficulty level of a multiple-choice test item is influenced by the position to which the correct choice has been assigned. They found that for five-choice items, those items having right answers in the fourth position are the most difficult; those with right answers in the second and third positions are the easiest; and the first and fifth positions are of equal moderate difficulty. Their results for four-choice items show that the third position is most difficult, and that difficulty

increases from the first through the third position, and decreases with the fourth. This finding is interpreted as agreeing with the results for five-choice items in that the next to last position was always found to be most difficult. Even though this position factor was relatively small, it was found to be statistically significant.

McNamara and Weitzman contend that an understanding of the material presented in a question is not the only factor at work in the selection of the correct answer. They believe that a subject does not always select an alternative on the basis of his fund of information alone, but is influenced to some degree by these positional factors. Essentially, their hypothesis is that since both the first and last choices in a list are more outstanding than the middle three (or two as the case may be), these inner choices become less noticeable and are less likely to be selected. Another possible explanation they offer is that a person going through the list without making a choice, is more likely to select the last choice rather than going through the list for a second time. This, however, is not sufficient to explain why the penultimate position was always found to be more difficult than the rest.

The present study tests the hypothesis that difficulty is a function of correct choice placement. Another concern of this study was to investigate the possible tendency among col-

¹ This report is based upon a master's thesis submitted to the University of Massachusetts in June 1958 and a paper presented at the annual convention of the American Psychological Association in September 1959.

Gratitude is extended to Jerome L. Myers for his critical review of the ideas presented here.

² Employed by the Electronic Systems Division in Bedford, Massachusetts; but now at the University of Massachusetts under Air Force training contract.

lege students to favor certain response positions to the neglect of others. For purposes of this investigation a response set was operationally defined as a deviation from chance at the 5% level of significance for the number of answers selected at each position.

METHOD

Subjects. Four hundred thirty-four students in the introductory psychology course at the University of Massachusetts served as subjects. Subjects were matched and divided into four groups on the basis of their test scores on the psychology midsemester examination. This was accomplished by randomly assigning students with similar test scores to one of the four groups.

Test. Items used in this study were selected from the previous year's final examination in psychology. Approximately 200 answer sheets from that examination were analyzed and a difficulty index, defined as the percentage of subjects passing an item, was computed for each of the 100 items on the test. Once these indices were arrived at the items were rearranged in order of increasing difficulty. This was done to create a uniform set throughout the test and to increase the probability of eliciting a position preference as the items become more difficult. To prevent item difficulty from being influenced by lack of time at the end of the test a time limit was imposed which was long enough to allow completion of all the items. Test items were all multiple-choice items and consisted of an incomplete statement which the examinee could complete correctly by selecting one of the four phrases following it. Below is a sample of the type of question used in the test.

Complex social needs are best related to:

- 1. primary incentives
- 2. secondary goals
- 3. complex human instincts
- 4. instrumental responses

Since the literature cites instances where positional response sets have occurred, it was recognized that not only may the student show a decided favoritism for certain positions, but the test constructor might also be affected by them in his placement of alternatives. Should the same positional factors manifest themselves in both the examinee's pattern of preferences and the test constructor's arrangement of alternatives, the probability of correct responses would be affected. To control for this bias an equal number of correct choices were randomly scattered among the four positions and four different forms of the test were developed so that the correct choice for each item appeared in a different position on each of the forms.

The position that each alternative was assigned for each item was determined by a scheme contrived by Mosier and Price (1945) for arranging and randomizing correct choices and distractors. All 24 permutations of the numbers one through four were

assigned positions in the test in accordance with the order of their appearance as determined by a table of random numbers. Alternatives for each item were rearranged to conform to the sequence number of the permutation assigned to it. The scheme was reapplied to determine the positions of alternatives for each of the remaining items. Each permutation was used once before any one was repeated. For the other three forms of the examination the entire cycle of alternatives was systematically rotated so that they appeared in a different position on each form. Each successive form was derived from the previous one. The matching of groups made it possible for the four forms of the test to be distributed equally among all levels of achievement.

Procedure. The subjects were instructed to read each item very carefully and then to mark on their answer sheet that alternative which they have decided is most correct. They were told not to omit any items and to guess when in doubt. Subjects were also informed that they would have sufficient time to complete all the items. Such a procedure provides a uniform response set thereby minimizing individual differences in responding resulting from the operation of other sets. All conditions for the collection of data for the four forms were identical. Papers of subjects who omitted items or who responded more than once to an item were discarded.

RESULTS

Analysis of variance was used to determine whether there were any differences between the means of the four test groups assigned different forms of the examination. The analysis reveals that the differences between means were not significant ($p > .20$) and that the forms were not altered in any way as to affect the test scores of the subjects.

When the total number of responses to each position (Table 1) were analyzed by means of the chi square test, for all forms combined, differences were found to be significant ($p < .05$). When the individual forms were analyzed in terms of the total number

TABLE 1
TOTAL NUMBER OF RESPONSES
TO EACH POSITION

Form	N	Position				Total
		1	2	3	4	
A	113	2,946	2,755	2,833	2,766	11,300
B	108	2,597	2,798	2,601	2,804	10,800
C	109	2,662	2,587	3,005	2,646	10,900
D	104	2,525	2,519	2,547	2,809	10,400
Total	434	10,730	10,659	10,986	11,025	43,400

TABLE 2
TOTAL NUMBER OF CORRECT ANSWERS
IN EACH POSITION

Form	N	Position				Total
		1	2	3	4	
A	113	1,791	1,856	1,727	1,760	7,134
B	108	1,664	1,694	1,749	1,704	6,811
C	109	1,693	1,723	1,826	1,860	7,102
D	104	1,717	1,568	1,640	1,681	6,606
Total	434	6,865	6,841	6,942	7,005	27,653

of responses to each position all were found to be significant at the .05 level also but on each form a different position was found to contribute most to the significance.

Application of the chi square technique to the total number of correct answers in each position (Table 2) failed to yield significance for all forms combined or for any of the individual forms with the exception of Form C, which was significant at the .05 level.

In order to determine whether the difficulty of any one position was significantly greater than any other, percentages of correct responses appearing in each position were computed (Table 3) both for all forms combined and for individual forms. This index was computed by dividing the total number of observed correct responses in each position by the total number of possible correct answers in each position.

Table 4 illustrates the percentage difference between the difficulty levels for the various positions and their critical ratios. The difference between percentages for positions two and four was the only one found to be

TABLE 3
PERCENTAGES OF CORRECT RESPONSES APPEARING IN
EACH POSITION ON EACH FORM

Form	Position			
	1	2	3	4
A	63.40	65.70	61.13	62.30
B	61.63	62.74	64.78	63.11
C	62.13	63.29	67.01	68.06
D	66.04	60.31	63.08	64.65
Total	63.27	63.05	63.98	64.65

TABLE 4
PERCENTAGE DIFFERENCES BETWEEN THE DIFFICULTY
LEVELS FOR THE VARIOUS POSITIONS AND THEIR
CORRESPONDING CRITICAL RATIOS

Correct response position	1	2	3
1			
2	.22 (.29)		
3	.71 (.96)	.93 (1.25)	
4	1.29 (1.69)	1.51 (2.04)	.58 (.79)

Note.—Critical ratios are in parentheses.

TABLE 5
PERCENTAGE DIFFERENCES BETWEEN THE DIFFICULTY
LEVELS FOR THE VARIOUS POSITIONS AND THEIR
CORRESPONDING CRITICAL RATIOS

Correct response position	1	2	3
1			
2	1.98 (1.35)		
3	.28 (.19)	2.26 (1.55)	
4	1.25 (.86)	.73 (.50)	1.53 (1.06)

Note.—Items over 50% difficulty. Critical ratios are in parentheses.

significant. Similar percentages were computed for all items in the test above the 50% level of difficulty. These 20 items were analyzed in an effort to heighten the effect of a position response set, since it has been shown by Cronbach (1946) that response sets are most apparent when items become more ambiguous or when they increase in level of difficulty. The percentage differences between the difficulty levels of the several positions for these items were not found to be statistically significant (Table 5). For these same 20 items chi squares were also computed for the total number of responses to each position and for total number of correct responses in each position. Significance was not obtained on either of these tests.

DISCUSSION

The research lends no support to the hypothesis that difficulty is a function of correct choice placement. Furthermore, a position preference hypothesis is untenable since the

position effect obtained was not consistently located. However, the findings do indicate that the difference between observed and expected frequencies for total number of responses to each position over all forms was significant. Although the fourth position contributed most to the significance it probably is of no great practical import since on each form the position favored was a different one. The fact that the so-called "position response set," as studied in the present investigation, does not carry over consistently from one form of the test to another suggests that there is no justification for referring to such phenomena as response sets. Certainly there was no consistent bias as implied by the McNamara and Weitzman data. The present findings are in accord with Cronbach.

To understand the lack of agreement in results among total responses on each form it is necessary to consider the effects of the location of very plausible distractors. Even when the position of all alternatives is determined randomly, as was done in this experiment, it is quite possible that the most plausible distractors within any one grouping of test items will not be evenly distributed among the four possible positions. Since the sequence of correct and incorrect answers was systematically varied from form to form this may more reasonably account for the differences in the distribution of total responses within forms. It is suggested that something other than a position response set is operating. Possibly the unequal attractiveness of distractors is what is contributing to the difference in frequency of response to each position as indicated by the systematic shifting of preferences to a different position from form to form. Another possible explanation might lie in the sequence effect of the position of the correct answer from item to item. Although there were an equal number of correct answers in each position throughout the test, the fact that an answer is in a certain

position on one item may influence the test taker to respond in a certain way on successive items. It should be understood that the arrangement of alternatives in a randomized fashion does not prevent an individual's response from being influenced by his previous responses. Actually the variance in results may be due to precisely such a set in the individual produced by the sequence effect.

The crucial question for an understanding of the problem of response set is the extent to which the set is stable or fixed. It is apparent from the data obtained in this study that the position effect is not constant since it shifts from form to form. Since response sets have no opportunity to show themselves when the subject gets most items correct they should be apparent, if they exist at all, on the more difficult items. Analysis of the data for items over the 50% level of difficulty revealed no significant differences whatsoever. In view of the nonsupporting results obtained in this study and the low reliability of position preferences reported by Cronbach it appears that position preference is not a significant source of invalidity in multiple-choice achievement tests. A more promising line of research might be to investigate the sequential effects from item to item or to attempt to get at a method for equalizing the attractiveness of distractors.

REFERENCES

- CRONBACH, L. J. Response sets and test validity. *Educ. psychol. Measmt.*, 1946, **6**, 475-493.
- CRONBACH, L. J. Further evidence on response sets and test design. *Educ. psychol. Measmt.*, 1950, **10**, 3-31.
- MCMANARA, W. J., & WEITZMAN, E. The effect of choice placement on the difficulty of multiple-choice questions. *J. educ. Psychol.*, 1945, **36**, 103-113.
- MOSIER, C. I., & PRICE, H. G. The arrangement of choices in m-c questions and a scheme for randomizing choices. *Educ. psychol. Measmt.*, 1945, **5**, 379-382.

(Received February 16, 1962)

INDICES OF EMPLOYER PREJUDICE TOWARD DISABLED APPLICANTS¹

THOMAS E. RICKARD,² H. C. TRIANDIS, AND C. H. PATTERSON

University of Illinois

A scale to measure prejudice toward disabled applicants for employment, based upon the multifactor stimuli method of Triandis, was developed. The scale was used with 2 samples, a group of 18 personnel directors and a group of 87 school administrators, who rated applicants for the position of accountant and 3rd grade school teacher. 6 types of disability, as well as absence of disability were included in the scale. All disabled groups were subject to expressed prejudice. The disabilities could be ranked in terms of the amount of prejudice expressed toward them. Competence and sociability were also influential in ratings, the former being more significant and complementary with disability. The method can be used to measure prejudice of various groups toward various disabilities in various settings.

It is generally accepted that unwarranted discriminations exist in the employment of the disabled (Federation Employment and Guidance Service, 1959; Jennings, 1951; Noland & Bakke, 1949; Schletzer, Dawis, England, & Lofquist, 1961). Since such discriminatory attitudes and practices are unwarranted in terms of the actual performance of the disabled (see, e.g., United States Department of Labor, 1948), they constitute prejudice. It was the purpose of this study to develop an instrument for measuring the extent of employer prejudice.

METHOD

Instrument

An assumption underlying the concept of prejudice toward the disabled is that where two persons apply for a job, and are equal in all characteristics except for the presence of a disability in one of them, the nondisabled person will be hired in preference to the disabled. In actual practice, such a situation seldom, if ever, exists. It is possible, however, to control or hold constant other variables than disability by the construction of a questionnaire or scale to which employers may respond.

The instrument is a modification of the questionnaire used by Triandis (1961) in a study of factors affecting employee selection in Greece and the United

States and developed by Triandis and Triandis (1960) in a study of social distance. The essence of this approach is that it describes the stimulus persons in terms of a number of characteristics simultaneously, rather than by the use of one characteristic as does the Bogardus Social Distance scale. Results are thus less ambiguous, since a particular characteristic can be measured while holding constant other characteristics. In the present study the instrument was developed to include four characteristics: disability (deaf, confined to a wheelchair, epileptic, discharged from a mental institution, discharged from prison, discharged from a tuberculosis sanatorium, and with no physical defect), sex (male or female), competence (barely competent, highly competent), and sociability (sociable, unsociable). Utilizing these characteristics in all possible combinations, 56 stimulus items were constructed. For each disability (and for no physical defect) the items were as follows, using deafness as an example:

Deaf, male, highly competent, unsociable
Deaf, female, highly competent, unsociable
Deaf, male, barely competent, sociable
Deaf, female, barely competent, sociable
Deaf, male, highly competent, sociable
Deaf, female, highly competent, sociable
Deaf, male, barely competent, unsociable
Deaf, female, barely competent, unsociable

The 56 items were randomized and presented with instructions requiring the subject to indicate his feeling about hiring each subject by circling a number on a seven-point scale from "I would *strongly recommend*" to "Would *strongly oppose*." Competence and sociability were defined. Judgments were to be made assuming a file on the applicant, but no personal interview.

This instrument may be used in relation to any job title. If the results are to be considered as relevant to prejudice, however, the position should be one which can be performed by persons with the

¹ From an unpublished dissertation entitled "Indices of Employer Prejudice: An Analysis of Psychological Aspects of Prejudice toward Disabled Workers," completed by the senior author at the University of Illinois for the PhD degree in the College of Education, January 1962.

² Deceased January 10, 1962.

disabilities included, that is, the disability is irrelevant to the performance of the position. The position of accountant was selected as one which would be common to a large number of work establishments, whose job requirements would be generally understood, and one which would also meet the requirement that it could be performed by a variety of disabled persons. The position of third grade teacher was used as a second position for one group of subjects to allow for a study of stability of rejection in positions with different requirements, and of stereotypes.

The instrument may be scored in a number of different ways. The primary interest in this study was employer prejudice, and therefore scores relevant to this were obtained. Two types of scores were obtained: prejudice scores of individual employers toward each disability, and prejudice scores for a group of employers. These are designated, respectively, as Index of Individual Employer Prejudice (IIEP) and Index of Group Employer Prejudice (IGEP). The individual scores for each disability are obtained by subtracting the sum of the eight ratings for a given disability (rejection score for the disability) from the sum of the ratings for the eight no physical disability items (rejection score for the nondisabled). The IGEP is the mean of the IIEP distribution for the group being studied. Both scores may thus be computed for any given disability and any given position.

Samples

Two widely different groups of potential employers of disabled persons were sampled. The first sample consisted of 32 personnel directors contacted by mail, of whom 25 returned the questionnaire, of which 18 were usable. The second sample consisted of 140 school administrators and potential administrators enrolled in classes in school administration at the University of Illinois, of whom 102 completed questionnaires, of which 87 were com-

pleted in full and were thus usable. The personnel directors completed the questionnaire only in regard to accountant, while the school administrators did so in regard to both accountant and third grade teacher.

Hypothesis

Employers, on the basis of preinterview knowledge, will reject disabled applicants more strongly than they will nondisabled; i.e., the Indexes of Employer Prejudice will be positive for all disabilities.

- 1. There will be differences among the six disabilities in the amount of employment prejudice expressed.
- 2. There will be some differences between the prejudice expressed toward the disabled who are being considered for the position of accountant and those who are being considered for the position of third grade teacher. Some disabled groups, such as the deaf, and the person in a wheelchair will experience more prejudice because of the relevance, or assumed relevance, of the disability to the job requirements. In other cases, such as the epileptic and the ex-prisoner, stereotypes are fixed, so that there will be little change of rejection with change of job title.
- 3. Competence, sex, and sociability, as well as disability, will affect employers' ratings.

RESULTS

Table 1 presents the Indexes of Group Employer Prejudice for both samples for accountant and for the school administrators for

TABLE 1
INDICES OF GROUP EMPLOYER PREJUDICE FOR SAMPLES OF PERSONNEL DIRECTORS AND SCHOOL ADMINISTRATORS

Disability	Personnel directors ^a		School administrators ^b			
	Accountant		Accountant		Third grade teacher	
	Mean	Variance	Mean	Variance	Mean	Variance
Epileptic	12.8	55.3	9.1	54.6	14.3	41.3
Prison	9.7	66.7	11.2	66.2	11.2	45.5
Mental hospital	9.0	66.7	8.5	55.4	11.1	45.9
Deaf	8.8	61.8	5.8	27.4	13.0	44.9
Wheelchair	7.4	61.2	3.7	21.5	7.9	36.4
Tuberculosis	2.8	15.8	3.3	17.0	4.9	22.9
Nondisabled	0.0		0.0		0.0	

^a N = 18.
^b N = 87.

TABLE 2

ANALYSES OF VARIANCE OF FACTORS INFLUENCING EMPLOYMENT RATINGS OF VARIOUS DISABILITY GROUPS BY PERSONNEL DIRECTORS

Source	Ex-tuberculars		Deaf		Wheelchair		Ex-mental patients		Epileptics		Prison parolees	
	SS	F	SS	F	SS	F	SS	F	SS	F	SS	F
Competence	50.5	405.4**	44.7	371.2**	43.5	357.1**	43.2	363.7**	35.6	316.0**	43.4	354.1**
Disability	.9	7.5**	7.4	61.3**	5.8	47.6**	7.6	64.3	17.3	153.8**	8.2	67.1**
Sex	.1	.5	.2	1.8	.1	1.2	.5	4.2*	.1	1.3	.3	2.4
Sociability	7.4	59.5**	6.0	50.09**	7.5	61.9**	7.1	60.1**	4.8	42.9**	5.4	44.1**
Interaction	3.1	2.3**	5.1	3.86**	6.0	4.5**	5.3	4.1**	7.9	6.4**	5.4	4.0**
Within cells	37.9		36.6		37.0		36.2		34.2		37.3	

Note.— $N = 18$.
 * $p < .05$.
 ** $p < .01$.

third grade teacher. The scores for each disability were significantly higher than the scores for the nondisabled for all three groups of comparisons ($p < .01$, one-tailed test) except for the difference between the tuberculous and the nondisabled for the personnel directors. The differences between the scores for accountant and teacher for the school administrator sample were significant ($p < .01$) for the deaf, the epileptic, and those in a wheelchair. In each case there was greater prejudice when the applicant was being considered for the position of teacher.

Table 2 shows the results of analyses of variance of the factors influencing the employment ratings of the personnel directors. The F for each disability is significant, indicating that disability was a significant factor in ratings. Ranking of the disabilities on the basis of the percentage of the total sum of squares attributable to disability yields the same order as that obtained from the IGEP scores. Competence was also significant in each case. Competence and disability are roughly complementary, that is, as the percentage of the total sum of squares due to disability increases, the percentage attributable to competence decreases.

DISCUSSION

The hypothesis that employers are prejudiced toward disabled applicants was supported. Disabled applicants were rejected more strongly than nondisabled applicants. The strength of the prejudice varies among the disabilities. Employers are more prejudiced toward the epileptic and persons dis-

charged from prison than toward the person discharged from a tuberculosis sanatorium. Persons discharged from a mental hospital, the deaf, and persons confined to a wheelchair fall in between. The results were similar for a sample of personnel directors and a sample of school administrators. The school administrators rejected the person discharged from a prison more strongly than the epileptic, but in general their prejudice appeared to be somewhat less than that of the personnel directors, except toward the tuberculous.

However, school administrators showed greater prejudice toward applicants for a position as third grade teacher than for applicants for a position of accountant, except in the case of persons discharged from a prison. The higher prejudice toward the deaf and persons in a wheelchair may be related to the relevance of the disability to the position. The greater prejudice toward the epileptic and the person discharged from a mental institution (the latter not significant, however) may indicate the presence of stereotypy in regard to these disabilities. However, to the extent that school administrators are concerned with the psychological adjustment of teachers the ratings of the ex-mental patients may not represent prejudice. The hypothesis that there would be differences between the prejudice expressed toward the disabled being considered for the position of accountant and those being considered for the position of third grade teacher related to apparent relevance of the disabilities and stereotypes, thus tends to be supported.

Hiring officers probably attempt to take into account the job requirements and the nature of the disability in making their decisions, although misinformation and stereotypy may still be present.

The hypothesis that competence, sex, and sociability would be significant factors in rating applicants was also supported, except for sex in the case of ex-tuberculars, persons confined to wheelchairs, the deaf, and epileptics. It appears that sex is not an important factor in qualifying for an accounting position for persons with these disabilities, while it is for persons who are prison parolees or ex-mental patients, with the females being at a disadvantage.

The F ratios for sociability were significant in all cases. It appears to be about as important as disability as indicated by the percentages of the total sums of squares in the case of persons in wheelchairs, the deaf, ex-mental patients, and parolees; and somewhat more important in the case of the tuberculous and less important in the case of the epileptics. These latter two are the cases in which disability has the least and the greatest weight, respectively, as a factor in the hiring decision. It is probable that the importance of the sociability factor would vary with different positions.

In every case, however, competence exerted the greatest influence in the hiring decision, contributing from 35.6% (for epileptics) to 50.5% (for the ex-tuberculars) to the total sums of squares. While competence was thus the major consideration in the decisions made by the personnel directors, the other factors were also significant. It appears that where the influence attributed to the disability is high, the variance is taken primarily from competence and secondarily from sociability.

In this study all variables (characteristics) were held constant for each disability group as a whole. This method allowed us to compare the influence of disability variables on the employers' decisions. It is possible to design studies to evaluate the influence of

the competence or sociability variable with the disability variable held constant. Under such conditions it is possible that we would find that an increase in the level of competence would offset the prejudicial effects of the disability. Such an investigation would experimentally examine a current dispute in the field of rehabilitation (Cruickshank, 1958, pp. 124-125; Patterson, 1962, p. 278). The present study demonstrates the importance of the competence factor in the employment of the disabled. It thus supports Patterson's idea that increased competence, perhaps obtained by so-called "overtraining," may compensate for the general prejudicial attitude of many employers toward the disabled applicant.

REFERENCES

- CRUICKSHANK, W. M. The exceptional child in the elementary and secondary schools. In W. M. Cruickshank and G. O. Johnson (Eds.), *Education of exceptional children and youth*. Englewood Cliffs, N. J.: Prentice-Hall, 1958.
- FEDERATION EMPLOYMENT AND GUIDANCE SERVICE. Survey of employer's practices and policies in the hiring of physically impaired workers. New York: FEGC, 1959. (Mimeo)
- JENNINGS, M. Twice handicapped. *Occup. Psychol.*, 1951, 30, 176-181.
- NOLAND, E. W., & BAKKE, E. W. *Workers wanted: A study of employers' hiring policies, preferences, and practices in New Haven and Charlotte*. New York: Harper, 1949.
- PATTERSON, C. H. *Counseling and guidance in schools: A first course*. New York: Harper, 1962.
- SCHLETZER, VERA M., DAWIS, R., ENGLAND, G. W., & LOFQUIST, L. H. Attitudinal barriers to employment. In *Minnesota studies in vocational rehabilitation*. No. 11. Minneapolis: Industrial Relations Center, 1961.
- TRIANDIS, H. C. Factors affecting employee selection in two cultures. In *Fourteenth International Congress of Applied Psychology*. Copenhagen: Munksgaard, 1961. Pp. 223-224. (Abstract)
- TRIANDIS, H. C., & TRIANDIS, LEIGH M. Race, social class, religion, and nationality as determinants of social distance. *J. abnorm. soc. Psychol.*, 1960, 61, 110-118.
- UNITED STATES DEPARTMENT OF LABOR. *The performance of physically impaired workers in manufacturing industries*. Washington, D. C.: United States Government Printing Office, 1948.

(Received February 27, 1962)

NEEDS, PERCEIVED NEED SATISFACTION OPPORTUNITIES, AND SATISFACTION WITH OCCUPATION¹

RAYMOND G. KUHLEN

Syracuse University

If major motives are satisfied in the context of work and career, then satisfaction with occupation should be a function of the discrepancy between personal needs and perceived potential of occupation for satisfying needs, particularly among those for whom occupation constitutes a major source of satisfaction (e.g., men rather than women), and in the instance of occupationally relevant needs, such as need achievement. The Edwards Personal Preference Schedule, a special rating scale, and a questionnaire were administered to 108 men and 95 women teachers. As predicted, discrepancy scores correlated .25 ($p < .01$) with occupational satisfaction for men, and .02 (ns) for women. Achievement need discrepancies were consistently related to occupational satisfaction. Other findings confirmed that occupation is psychologically more central for men.

It may be hypothesized that satisfaction or dissatisfaction *with* an area of life is a function of the degree to which one finds satisfaction for major needs *in* that area of living. This presumably will hold true especially among those for whom a given area of living (e.g., occupation) represents a major source of life satisfaction, and may not obtain at all in a sphere of life with which a person is little concerned, even though he is participating. Schaffer (1953) has advanced a similar hypothesis with respect to work (though without the latter restriction) and has presented supporting evidence. In the area of marriage Ort (1950) has reported a correlation of $-.83$ between happiness and frequency with which role expectations (needs) were not satisfied.

In the present investigation, a first hypothesis was that those individuals whose measured needs are relatively stronger than the potential of the occupation for satisfying those needs (as they perceived this potential) will tend to be frustrated and hence to be less well satisfied with their occupation. Where needs and the perceived need-satisfaction potential of the occupation are more in harmony, it was anticipated that satisfaction with occupation would be rated higher. However, since a career role tends to be primary

for males and relatively secondary for females, a second hypothesis was that these relationships will hold to a greater degree among men than among women. Relevant to this hypothesis is the finding of Brayfield, Wells, and Strate (1957) that job adjustment is correlated with general adjustment to a higher degree among men than among women, a finding which they attributed to the greater importance of work in the life scheme of men.

The general hypothesis of this phase of the study relates to the degree of overall satisfaction or frustration of needs experienced in the occupation. The specific needs involved are, for this hypothesis, unimportant, as long as they are vocationally relevant. One man, with high achievement needs, is frustrated because he sees no future. Another, with strong dominance needs, is irked by the submissiveness required of him. Both are in the same occupation; both are frustrated and dissatisfied. It is recognized that certain needs may not be perceived as being satisfiable in the occupational context whereas others are. Thus the satisfaction or frustration of the need for achievement would presumably bear a relationship to satisfaction with career, whereas sex needs typically would not. In fact, it may be assumed that career is a major source of satisfaction for the achievement need and thus it was predicted (a third hypothesis) that satisfaction of this need would be particularly important for (i.e., more highly related to) occupational satisfaction.

¹ This study was done pursuant to a contract with the Cooperative Research Branch of the Office of Education, United States Department of Health, Education, and Welfare.

METHOD

Subjects

Students in certain of the writer's graduate classes, enrolling mainly teachers-in-service, were tested in the present phase of the investigation. Of some 323 tested, complete data were available for 203 (108 men and 95 women) who were engaged in junior and/or senior high school teaching. These 203 people constituted the major sample, though numbers varied slightly from analysis to analysis (in some instances more, others less than 203) since it was desired to capitalize all the data available. The subjects were mainly in their 20s and 30s.

Procedure

The subjects were asked to respond to three instruments, administered in the following order: the Edwards Personal Preference Schedule (a measure of needs); a questionnaire which asked for ratings of satisfaction with present job and occupation, and for other information relating to job satisfaction and plans; and an instrument entitled "Personality Types and Occupations" which was designed to obtain estimates of the perceptions the respondents had of the need-satisfaction potential of their occupation. In some instances these instruments were administered in the same sitting, but mainly the data were collected in two separate sessions with the Edwards scale constituting the first session.

Rating of satisfaction with occupation was on an 11-point scale, with instructions to "think of your occupation in general, not your particular job." Similar ratings were obtained with respect to satisfaction with present position. Previous research (Johnson, 1955) has shown that ratings so obtained correlate .64 with job satisfaction scores obtained from an extensive questionnaire, and .61 with pooled ratings of teachers by colleagues as to their job satisfaction, the latter sample, however, being quite small ($N=18$). Test-retest reliability over 3 weeks was .89. The questionnaire also contained questions inquiring as to whether or not the occupation was something they truly wanted to do or in which they planned to continue. Strong (1955, pp. 98-117) had found answers to these questions related to other evidences of satisfaction (see Table 1).

Personality Types and Occupations was the same instrument employed in a previous study (Kuhlen & Dipboye, 1959) with slight modification of directions to make it appropriate for people already employed. "Personality types," which were actually the descriptions (or slight modifications thereof) of various needs from Edwards' (1954) test manual, were presented with instructions for the subject to rate the degree to which a person of this type would likely be satisfied or frustrated in the teaching profession.²

² Heterosexual needs, though included in Edwards scale, was not included in the present scale because of the disruptive influence occasioned by the humorous reaction generated by a trial form.

In assigning such ratings he was instructed to ignore whether the type being considered would make a "good" member of this profession, to ignore his own attitude toward this type of person, and to think of the occupation in general, not a particular school or system. An example, relating to achievement need, will illustrate this device.

Type 1. This person has a high need to achieve. He likes to do his best to accomplish tasks requiring skill and effort, to be a recognized authority, to accomplish something of a great significance, to do a difficult job well, to solve difficult problems, to be able to do things better than others, to get ahead, to be a big success.

Will this occupation offer him opportunity for satisfying experiences or will it pose frustrations? And to what degree? Circle one number to indicate your judgment.

-5 -4 -3 -2 -1 0 +1 +2 +3 +4 +5

This occupation will pose exceedingly high frustration	No special satisfactions or frustrations	This occupation will offer exceedingly high satisfactions
--	--	---

One focus of the present study is upon the degree to which satisfaction with occupation (in this instance, teaching) is related to the discrepancy between strength of one's basic need and the perceived potential of the occupation for satisfying those needs. Discrepancies were determined in the following fashion: First, index values of 1 through 5 were assigned to indicate on the same scale (a) strength of needs and (b) perceived potential of occupation for satisfying those needs. Edwards raw scores were converted into standard scores (based on norms) before index values were assigned. The index values had the following meanings:

Index value	Edwards score	Need-satisfaction potential
1	-34	-5, -4
2	35-44	-3, -2
3	45-54	-1, 0, +1
4	55-64	+2, +3
5	65+	+4, +5

Discrepancies were computed by subtracting the index assigned to need strength from the index assigned to the perceived need-satisfaction potential of the occupation. Thus a person with need-strength of 4 (between .5 and 1.5 *SDs* above the mean on Edwards norms) on achievement need who perceived need-satisfaction potential of the occupation as being at an index value of 2 (i.e., rated it -2 or -3 on the scale) would have a "need-need satisfaction discrepancy" of -2. Negative discrepancies thus identify instances of presumed probable frustration, whereas zero or positive discrepancies imply adequate or more than adequate opportunities for satisfaction in the occupation, as perceived by the particular individual.

It was predicted that negative discrepancies would be associated with low satisfaction with occupation, particularly in the instance of needs (e.g., need for achievement) which are commonly satisfied through occupation and career. In general, it was anticipated that positive discrepancies (implying ample opportunity for need satisfaction) would be associated with high satisfaction in occupation. However, it is conceivable that an occupation that offers considerable opportunity for the satisfaction of a particular need may actually be frustrating to a person low in this need, if, along with the opportunity, colleagues or superiors expect or demand a high level of motivation of the particular type. It was not anticipated that the occupation of public school teaching would be "demanding" with respect to the type of needs here studied and analyses were not designed to reveal curvilinear relations between satisfaction with occupation and the need-need satisfaction discrepancies.

SATISFACTION WITH OCCUPATION

Generally speaking, this group of subjects was quite satisfied with their careers. Their occupational satisfaction ratings ranged from 1 to 11 for 108 males and from 3 to 11 for 95 females, with the respective medians being 8.9 and 8.8. About three-fourths of the ratings of each sex fell in the categories of 8, 9, and

10. A rating of 6 represented "average satisfaction."

Table 1 presents the distributions of answers to questions which might also be expected to reflect job satisfaction. It will be noted that the vast majority of both sexes (a) felt that teaching was something they truly wanted to do, or approximately so, and (b) wanted to continue in or at least had no plan to leave.³ (However, in this sample the majority of men indicated they were contemplating changing or making an effort to change their particular *positions*.) The mean occupational satisfaction-ratings of those selecting various alternatives in this table suggest that the questions and ratings are measuring the same variable, to a degree at least.

It is to be noted, in line with the hypothesis that relationships will be less pronounced in a group for whom career has lower saliency, that there is a reliable relationship between

³ In this connection, it is of interest that only nine men and five women rated their positions as less satisfactory than the typical teaching position.

TABLE 1
OCCUPATIONAL SATISFACTION OF THE TEACHER SAMPLE

		Men			Women	
		Mean occupa- tional satisfac- tion	Mean job satisfac- tion		Mean occupa- tional satisfac- tion	Mean job satis- faction
		<i>N</i>				
Attitude toward career						
Truly wanted to do	39	9.59	9.28	25	9.40	8.80
Approximately what wanted	49	8.47	7.90	47	8.74	7.96
Came to accept	9	7.50	6.75	5	6.84	7.79
Would not have chosen	10			14		
Dislike	1			0		
<i>F</i>		13.69***	11.46***		25.29***	1.93
Permanence of career						
Want to continue	53	9.17	8.85	40	9.08	8.60
No plan to change	38	8.66	8.13	34	8.47	7.68
Contemplate change	7	7.29	6.24	9	7.44	8.12
Making effort to change	3			2		
Will definitely change	7			5		
<i>F</i>		9.45***	11.38***		7.61***	1.40

*** *p* < .01.

TABLE 2
MEAN SCORES OF THE TEACHER SAMPLE

Edwards PPS need subscales	Edwards score				Ratings of perception of need- satisfaction potential ^a			
	Males		Females		Males		Females	
	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>
Ach	50.1	9.9	50.5	10.5	0.5	3.3	0.2	3.5
Def	52.9	9.3	53.1	11.0	0.7	2.8	0.5	3.2
Ord	51.1	11.8	52.1	11.7	1.8	2.8	1.2	3.2
Exh	49.3	10.4	49.7	11.3	-0.1	3.4	0.3	3.4
Aut	49.8	9.6	47.6	9.3	-2.9	2.4	-3.2	2.4
Aff	49.1	9.9	48.0	10.6	2.5	2.0	2.2	2.2
Int	49.6	10.4	51.6	9.8	3.4	1.8	2.9	2.2
Suc	48.0	9.9	46.8	10.6	-2.7	2.1	-2.6	2.4
Dom	48.5	9.8	47.6	9.3	1.7	2.8	1.0	3.0
Aba	50.0	10.0	48.8	9.9	-2.4	2.6	-3.1	2.3
Nur	50.5	9.5	49.4	10.2	3.0	1.9	3.6	1.9
Cha	49.0	11.2	50.7	10.5	-0.1	3.5	0.2	3.3
End	52.9	10.5	52.6	11.1	2.6	2.3	2.3	2.7
Het	47.6	10.5	51.4	10.8	—	—	—	—
Agg	51.7	9.3	51.0	9.7	-2.7	2.6	-3.3	2.5
<i>N</i>	107		91		107		91	

^a On this rating, +5 means this occupation will be extremely satisfying to a person high in this need; -5 indicates the likelihood of extreme frustration of such a person.

satisfaction with *current* position and statements reflecting attitude toward career choice and expected permanency of career in the case of men, but not in the case of women. The correlation between *ratings* of satisfaction with current position and with occupation were also higher in the case of men than women, the respective *r*'s being .62 and .36.

NEEDS, PERCEPTIONS, AND OCCUPATIONAL SATISFACTION

Table 2 contains the means and *SD*s for the two sexes for each of the 15 needs and for ratings as to the potential of the occupation for satisfying those needs. The Edwards raw scores were translated into standard scores according to the published norms, which have a mean of 50 and an *SD* of 10. Thus one can make certain comparisons of teachers with Edwards norms, if he so desires, though age and probably marital status differences make such comparisons tenuous. Compared to the norms, teachers are at about the mean in achievement need, are high in deference, low

in succorance (women), high in endurance, to select examples.

The instrument devised to obtain ratings of the perceived potential of the occupation for satisfying various needs yielded especially pertinent information. Table 2 presents mean values (ratings were on an 11-point scale, -5 to +5) for each of 14 needs. Both sexes agreed that needs for affiliation, intraception, dominance, nurturance, and endurance might be readily satisfied in the teaching profession, but that the individuals with strong needs for autonomy, succorance, abasement, and aggression would likely be extremely frustrated. These findings would be anticipated through even a casual evaluation of the teacher's role and activities. In fact, the latter needs might be expected to be frustrated in most work situations.

However, an unanticipated finding was the marked differences in perception of the need-satisfaction potential of teaching with respect to certain needs. The overall distributions of ratings are presented in Table 3 for the three

TABLE 3
DISTRIBUTION OF RATINGS OF PERCEIVED NEED SATISFACTION OF TEACHING

Need-satisfaction potential ratings	Type of person rated					
	High achievement need		High exhibition need		High change need	
	Male	Female	Male	Female	Male	Female
Highly frustrating						
-5	11	14	13	12	17	12
-4	8	5	8	6	10	6
-3	13	13	14	10	14	6
-2	9	5	10	6	3	10
-1	9	3	6	3	3	4
0	0	6	6	12	7	9
+1	6	6	12	6	10	8
+2	14	9	8	9	13	8
+3	17	18	11	13	15	16
+4	17	6	10	9	12	11
+5	9	13	14	12	9	6
Highly satisfying						
Number of ratings ^a	113	98	112	98	113	96

^a These *N*s differ slightly from those reported in the preceding table. All of the data was used here, whereas the previous table is based on cases with complete data.

needs (achievement, exhibition, and change) having the largest *SD*s.

The facts for achievement need are especially interesting. It had been anticipated that teachers would view teaching as being somewhat frustrating to individuals with strong achievement needs. But the *mean* rating suggested only that the profession would be neither especially frustrating nor satisfying. Actually the distribution of ratings is sharply bimodal. Many view teaching as extremely satisfying to the high achievement need person; many others view teaching as highly frustrating to such a person. Relatively few view it as in between. And this is true for both sexes. One can only speculate regarding the reasons for sharply divergent views. Is one group perceptive as to avenues for advancement in public education, the other not? Is one group from low socioeconomic and occupational background, and thus views the teaching profession as evidence of marked achievement, whereas the other group represents offspring of high socioeconomic or occupational level parents?

A similar tendency toward bimodality is evident in the case of "exhibition." Perhaps

one group perceives the opportunities for the satisfaction of need exhibition before students, and the other is more aware of the unacceptability of such behavior on the part of a teacher in the community. In the instance of the need for change, it may be that the teachers who see opportunities for satisfaction of this need are in progressive schools while those who view teaching as frustrating to the need for change are in static schools. To be sure, though, certain kinds of people may perceive a static profession as frustrating to the need for change whereas others perceive it as offering unusual opportunities for change precisely *because* of its long-term static character.

In any event, these findings illustrate that perceptions of an occupation may vary greatly from one person to another. The need for further study of the variables that produce these contrasting perceptions is apparent.

The next analysis involved the computation of correlations between measured needs and the ratings of the need-satisfaction potential of the occupation, and between these two variables and rated occupational satisfaction. Table 4 contains the results. It will be

noted that, in the data for men, reliable correlations existed between the measured need and rated need-satisfaction potential of the occupation for about half the needs, the highest r being .27 in the instance of dominance. Throughout the table one notes fewer reliable correlations for women.

Although Table 2 suggested that teachers as a group were average in achievement need and saw no special frustration of this need in the teaching profession, the facts in Table 4 which involve rated satisfaction as related to the raters' needs are in line with the prediction. In the case of men, those with high achievement needs tend to be dissatisfied, but those who *perceive* teaching as being potentially satisfying to the high achievement need person tend to be satisfied. In the case of women, only the latter was true. Among men, again, the data suggest that high autonomy need individuals are likely to be frustrated, and that those with high dominance needs are likely to be satisfied. Those who perceive the teaching as potentially satisfying to the high endurance need person tended to be more

satisfied. Among women, only one of the 29 correlations was reliable, suggesting that for them satisfaction with occupation is not so dependent upon need satisfaction as is true of men. This finding is in line with the hypothesis that such relationships will be lower among those for whom occupation has low saliency.

It will be recalled that other measures of occupational satisfaction were available in addition to the overall rating. Two questions asked that the subjects categorize themselves with respect to their attitudes toward their career and their judgments as to their permanence in the profession (see Table 1). Three groups were set up with respect to each question: those who checked Response Number 1 constituted one group; Response Number 2, the second group; and those who checked either Response Numbers 3, 4, or 5 constituted the third group. The three groups presumably varied in degree of satisfaction with the third group being least satisfied. Separate analyses of variance (simple one-way classification) were computed for each need and for

TABLE 4
CORRELATIONS AMONG MEASURED NEEDS AND RATINGS

Need	Correlations between:					
	Need and perception		Satisfaction and need		Satisfaction and perception	
	Male	Female	Male	Female	Male	Female
Ach	.03	.10	-.19*	-.02	.21**	.32***
Def	.11	.14	.07	.11	-.12	-.05
Ord	.00	.16	.04	.01	-.06	-.10
Exh	.19*	.14	-.01	.08	.08	-.19
Aut	.25***	.09	-.28***	-.19	.04	.06
Aff	.15	.22*	.04	.00	-.03	.16
Int	.20*	.05	.05	.03	.03	.14
Suc	.21*	.20	-.07	-.03	-.05	.01
Dom	.27***	.06	.20*	-.12	.22*	-.05
Aba	.23*	.08	.15	-.03	-.10	-.20
Nur	.12	.10	.14	-.01	.12	-.16
Cha	.24*	-.01	-.12	.05	.16	.03
End	.05	.27**	-.11	.19	.20*	.03
Het	—	—	-.04	-.01	—	—
Agg	.11	.13	.14	-.10	.17	-.11
<i>N</i>	107	91	107	91	107	91

Note.—All levels of confidence represent two-tailed tests, except in the case of achievement need where a one-tailed test was justified by a specific directional prediction.

* $p < .05$.

** $p \leq .025$.

*** $p \leq .01$.

TABLE 5

ANALYSIS OF VARIANCE SHOWING RELATIONSHIP BETWEEN NEED FOR ACHIEVEMENT AND PERCEPTION OF THE OCCUPATIONS POTENTIAL FOR SATISFYING ACHIEVEMENT NEEDS VERSUS ATTITUDE TOWARD CAREER AND EXPECTED PERMANENCE OF CAREER

	Grouped by attitude toward career				Grouped by expected permanence of occupation			
	1 High	2 Medium	3 Low	F	1 High	2 Medium	3 Low	F
Need ^a								
Men	48.8	51.6	48.9	0.82	49.5	49.4	53.5	1.65
Women	52.0	48.1	55.3	3.19*	51.6	49.7	49.9	1
Perception ^a								
Men	6.3	7.1	4.0	3.50*	6.8	6.5	5.1	2.16
Women	7.3	5.9	5.2	2.34	6.1	6.5	5.9	1
Number of subjects								
Men	39	49	20		53	38	17	
Women	25	47	19		40	34	16	

Note.—For meaning of the groups see Table 1.
■ Achievement.
* $p < .05$.

each rating as to need-satisfaction potential, with the sexes separate.

This analysis was not particularly informative. Only 9 of 116 comparisons yielded reliable differences. The three that occurred (out of 58) in the case of women is the number expected by chance. These included *achievement* need in the instance of attitude toward career and perception of occupation with respect to *abasement* and *aggression* in the instance of expected permanence of career. Six were reliable among the 58 comparisons involving men. But again need for achievement seemed to be a significant variable in work satisfaction. These included need for *change* in the instance of attitude toward career and in the same classifications perception of occupation with respect to *achievement*, *affiliation*, and *dominance*, and in the classifications related to attitude toward career, perception with respect to *autonomy* and *change*. As Table 5 shows, in the four comparisons involving need achievement and attitude toward career (the two sexes, need and perception), two were characterized by significant *F*s. The women who were dissatisfied had highest achievement needs, and those (both sexes, though only one reliably) who were most satisfied tended to *perceive* the occupation as potentially satisfying to the high achievement need person. No significant dif-

ference with respect to achievement need appeared when the subjects were classified according to attitude toward career.

DISCREPANCIES BETWEEN NEEDS AND NEED-SATISFACTION POTENTIAL

Although several tables of findings have already been presented the main focus of the study was the relationship between occupational satisfaction and the *discrepancy* between strength of one's needs and the perceived potential of the occupation for satisfying the particular needs.

An overall index of the degree to which the array of needs studied was viewed as being satisfiable or susceptible to frustration in the occupational context of public school teaching was computed *for each individual* by summing algebraically the discrepancies for the 14 needs (see above for description of the discrepancy score). This total index was then correlated with satisfaction-with-occupation ratings. The obtained correlations for the teacher sample of 108 men and 95 women were .25 and .02, respectively. The value for men is significantly different from zero at the .01 level of confidence (one-tailed test) while that for women obviously is not. The two correlations differ reliably at the .01 level of confidence.

This finding may be interpreted as supporting in the case of men the general hypothesis that satisfaction with occupation is a function of the degree to which one's array of needs can be satisfied in that occupation. It should be noted that the procedure employed did not ask directly the degree to which it was expected that particular needs would be satisfied or were actually satisfied *in the occupation*. One would expect higher correlations between such an index, if it could be obtained, and satisfaction with occupation than were obtained with the present index, the obtained correlation probably being somewhat attenuated by lumping together occupationally relevant and irrelevant needs. As anticipated the correlation for women was smaller than that for men, though it had been expected that this correlation also would be reliably positive. The low correlations, even for the men, may also be attributed to the probability (suggested in a previous study by Kuhlen & Dipboye, 1959) that teachers as a group are *not* career-minded, i.e., are not the type of people who look to career as a major source of satisfaction.

To test the relationship of discrepancies between a specific need and the perceived po-

tential of the occupation for satisfying that need, the data were dichotomized so that those individuals whose need strength equaled or was less than the potential for satisfying that need were in one group (0 or + discrepancies) and the remainder (− discrepancies) were in the second group. Occupational satisfaction ratings were dichotomized roughly at the median, with those with ratings of 9 and above falling in the top groups, numbering 71 men and 56 women. The dissatisfied had ratings of 8 or below and numbered 37 men and 39 women. The data were then classified in 2 × 2 tables and the null hypothesis tested by chi square. Table 6 shows the findings for those three needs where significant differences occurred for at least one of the sexes. It will be noted that the prediction with respect to the achievement need is supported .05 level in the case of women, but not in the case of men (*p* in this instance, < .10 > .05). Thus the finding gives only tenuous support to the hypothesis, though the differences were in the same direction for both sexes. It had been anticipated that the stronger relationship would hold for men.

Though no specific hypotheses were formulated with respect to endurance and exhibi-

TABLE 6
PERCENTAGE OF HIGH AND LOW OCCUPATIONAL SATISFACTION GROUPS WITH FAVORABLE
"DISCREPANCY" SCORES

Need	Low satisfaction (%)	High satisfaction (%)	Chi square	<i>p</i>
Ach				
Men	49	68	2.92	<.10>.05
Women	48	73	4.92	<.05>.02
Exh				
Men	57	60	.013	ns
Women	77	50	5.92	<.02>.01
End				
Men	68	89	5.89	<.02>.01
Women	72	88	2.74	<.10>.05
<i>N</i>				
Men	37	71		
Women	39	56		

Note.—This table may be read as follows: 49% of the men who fell in the low satisfaction group had, in the case of achievement need, discrepancy scores indicating that their rating of the need-satisfaction potential of the occupation was equal to or greater than their need strength (discrepancy scores of 0 or +).

tion, the finding is reasonable in the former. But the excess of zero or positive discrepancy scores in the *low satisfaction* group of women in the instance of the need for exhibition is hard to explain. A possible explanation is that those who see the job as a place for gaining satisfaction of this need run into other difficulties that lower their satisfactions. But it should be noted that only 5 of 58 analyses yielded significant *F*s, and thus focusing upon these differences runs considerable risk of capitalizing chance. This would not be the case in the instance of achievement need, however, since a specific prediction was made in this instance.

DISCUSSION

The findings tend to support the general hypotheses of the study in the case of men, with respect to the array of needs, and with particular reference to the achievement need regarding which a specific prediction was made. While those correlations which were significantly different from zero were low in the case of men, correlations were generally lower for women, with only an occasional *r* reaching significance. Both the low correlations for men and the less positive findings for women are reasonable in view of the fact that as already noted, a major restriction must be placed on the hypothesis that occupational satisfaction is a function of the degree to which needs are satisfied in the occupation. This hypothesis would be expected to hold only for the people, and for occupations which attract the kind of people, who view occupation and career as a major source of need gratification; i.e., people for whom career and occupation have high "saliency." Other evidence (Kuhlen & Dipboye, 1959) suggests that career is less salient for teachers than for other occupational groups. Also occupation appears to be clearly a secondary role for women (i.e., not a primary source of need gratification), especially for *young* single teachers and for married teachers (Kuhlen & Johnson, 1952).

Although the present study was conducted with subjects homogeneous with respect to occupation, it would be expected that the major hypothesis relating to the array of needs would also be supported, and probably

more clearly, in an occupationally heterogeneous sample. The major hypothesis does not and need not, specify particular needs; but only that occupationally relevant "needs" in general must be satisfied in the occupation if satisfaction is to be found. Findings which relate to the relationship of frustration or satisfaction of *particular* needs might be expected to be relatively specific to the occupation, and presumably would be suggestive of potential tension areas, worthy of attention of workers, of those selecting personnel, of supervisors in that occupation, or of those concerned with assisting young people to sound vocational decisions. In certain situations, for example, it is probably undesirable to employ people with particular need patterns. For many jobs, the ambitious man, the aggressive, dominant go-getter, may be an extremely poor choice. And not infrequently there is a conflict between the type of person needed to do a particular job and the potential of the position for satisfying fundamental career needs of the person who can do that job.

REFERENCES

- BRAYFIELD, A. H., WELLS, R. V., & STRATE, M. W. Interrelationships among measures of job satisfaction and general satisfaction. *J. appl. Psychol.* 1957, **41**, 201-208.
- EDWARDS, A. L. *Edwards Personal Preference Schedule (manual)*. New York: Psychological Corporation, 1954.
- JOHNSON, G. H. An instrument for the measurement of job satisfaction. *Personnel Psychol.* 1955, **8**, 27-38.
- KUHLEN, R. G., & DIPBOYE, W. J. Motivational and personality factors in the selection of elementary and secondary school teaching as a career. Technical report, 1959, United States Office of Education Cooperative Research Program, Washington, D. C.
- KUHLEN, R. G., & JOHNSON, G. H. Changes in goals with increasing adult age. *J. consult. Psychol.*, 1952, **16**, 1-4.
- ORT, R. S. A study of role-conflicts as related to happiness in marriage. *J. abnorm. soc. Psychol.*, 1950, **45**, 691-699.
- SCHAFER, R. H. Job satisfaction as related to need satisfaction in work. *Psychol. Monogr.*, 1953, **67**(14, Whole No. 364).
- STRONG, E. K., JR. *Vocational interests 18 years after college*. Minneapolis: Univer. Minnesota Press, 1955.

(Received February 27, 1962)

NIGHT VISION SENSITIVITY DURING PROLONGED RESTRICTION FROM SUNLIGHT

JO ANN S. KINNEY ¹

United States Naval Medical Research Laboratory, Groton, Connecticut

The night vision sensitivity of a group of 24 men was tested monthly during the 3-month submerged cruise of the Triton. There was no evidence in the test scores that night vision sensitivity can be improved beyond its seasonal peak by further restriction from sunlight.

A previous study (Sweeney, Kinney, & Ryan, 1960) has shown that night vision sensitivity varies seasonally with amount of exposure to sunlight. For individuals in northern temperate climates, night vision sensitivity is poorest in the summer and becomes progressively better during the fall and winter. This result suggested that sensitivity might be further increased if individuals were cut off from sunlight completely, an extreme which is not met even in midwinter. The course of night vision sensitivity therefore was measured on a group of men aboard the USS Triton, SS(n)-586, during its 3-month submerged trip around the world.

PROCEDURE

The NMRL Night Vision Sensitivity Test was used (Kinney, Sweeney, & Ryan, 1960; Sweeney, Kinney, & Ryan, 1959). The apparatus presents small spots of light of varying size throughout the visual field and provides a score for each subject composed of the total number he correctly identifies.

Twenty-four men served as subjects while aboard the submarine. They were tested in late February, shortly after the Triton left port, and in the latter parts of March and of April. A final test was made during June and early July after the Triton had returned to port and the men had returned from leave.

Before testing began, the subjects wore red goggles for 15 minutes and dark-adapted for an additional 10 minutes. The test was then given, a brief rest period followed, and the test was repeated. There are, therefore, eight test scores for each subject, two each in February, March, April, and June.

¹ The author wishes to thank James E. Stark, Commander, Medical Corps, United States Navy, who directed the testing during the cruise; and the many other individuals aboard the Triton whose cooperation and assistance made the study possible.

RESULTS

Table 1 presents the average test scores, standard deviations, and test-retest correlations for the 24 men at the different testing periods. There is no evidence of a gain in scotopic sensitivity under the submerged conditions of the Triton cruise. The small gain in test scores from February to March, as well as the gains found between the first and second tests in a single day, are readily explained

TABLE 1
RESULTS OF 24 MEN ON NIGHT VISION SENSITIVITY TEST

Month	Test	Retest	<i>M</i>	<i>t</i> ^a
February				
Score	60.2	66.7	63.4	—
<i>SD</i>	12.1	9.9	10.5	—
<i>r</i>	.83			
March				
Score	63.8	67.2	65.5	—1.18
<i>SD</i>	7.8	9.0	8.4	—
<i>r</i>	.93			
April				
Score	62.4	63.5	63.0	.24
<i>SD</i>	10.7	11.0	10.6	—
<i>r</i>	.92			
June				
Score	58.0	59.6	58.8	2.40
<i>SD</i>	11.8	12.2	11.9	—*
<i>r</i>	.88			

^a Calculated between scores in February and the month in question.
* *p* < .05.

by a small practice effect always found in repeated testings.

The decrease in test scores found in June, as compared with the other months, shows once again the seasonal effect of exposure to sunlight. The difference is statistically significant and similar in magnitude to that found previously.

DISCUSSION

Since the Triton left port in midwinter, the night vision of the men may be assumed to be at its best level for normal seasonal living. The scores obtained by the men during their 3 months of restriction from sunlight did not give any indication that sensitivity may be increased beyond its peak winter level. This does not preclude the possibility that sensitivity could be increased under conditions of more severe light restriction, although one would expect to find a trend under the conditions of this study if a large shift in sensitivity were basically possible. The amount of restriction from light was considerable considering the fact that light levels of 1,000 to 10,000 foot-lamberts regularly would be encountered by individuals in a normal environment. The light levels within the submarine, however, remained relatively stable at 10–20 foot-lamberts or less.

The expected result from repeated testings under constant light conditions after sensitivity has reached its peak is a slight increase in test scores through practice. This was found in the average scores for March but not for April. The decrease, although small, is unexpected since repeated use of the night vision test has never yielded a group mean for a retest that was poorer than the mean of the original test.

Further analysis of the data revealed normal distributions of differences between test scores in the February, March, and April tests; but showed a bimodal distribution between the April and June tests. The majority of the men showed a large decrease in score in June as would be expected from the sudden exposure to high summer light levels. Seven of the men, however, had relatively poor scores in April, followed by a large increase in June. Figure 1 shows the test

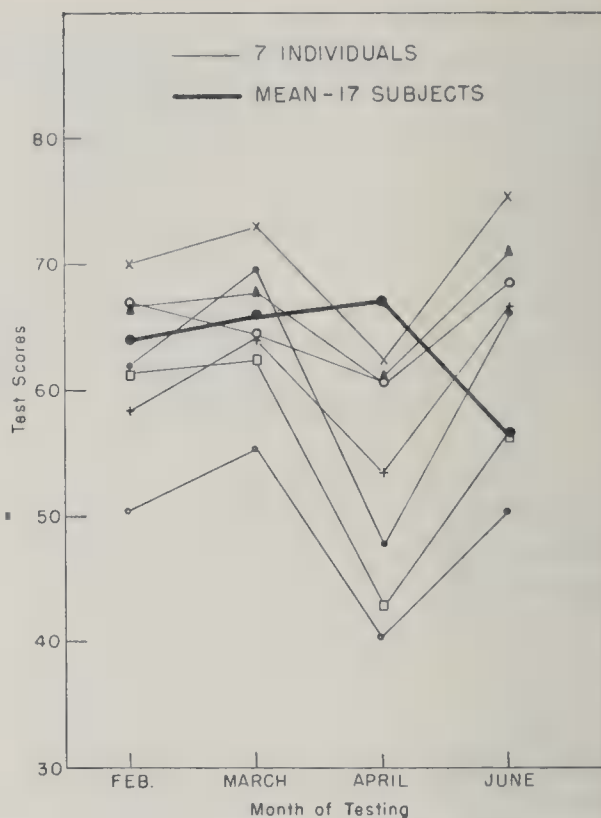


FIG. 1. Test scores for 7 individuals and for mean of 17 other subjects.

profiles of these 7 men together with the mean of the other 17.

Data on three of these seven men were available from an attitude study that was run concurrently during the Triton cruise (Weybrew, 1962). Their profiles on a morale factor revealed a severe loss in morale during the latter part of April as compared with the latter part of March. This pattern of morale loss was not reflected in the attitude data for the group. Granger (1957), in his review of the literature on the relation between night vision and psychiatric disorders, has shown that loss in sensitivity is often correlated with various personality deviations. Obviously nothing definitive can be stated with the small group here; the data are given as an interesting possibility of another connection between personality factors and sensitivity measures.

REFERENCES

- GRANGER, G. W. Night vision and psychiatric disorders: A review of experimental studies. *J. ment. Sci.*, 1957, 103, 48–79.

- KINNEY, JO ANN S., SWEENEY, E. J., & RYAN, ALMA P. A new test of scotopic sensitivity. *Amer. J. Psychol.*, 1960, **73**, 461-467.
- SWEENEY, E. J., KINNEY, JO ANN S., & RYAN, ALMA P. Standardization of a scotopic sensitivity test. *U. S. Naval Med. Res. Lab. Rep.*, 1959 (Mar), No. 308.
- SWEENEY, E. J., KINNEY, JO ANN S., & RYAN, ALMA P. Seasonal changes in scotopic sensitivity. *J. Opt. Soc. Amer.*, 1960, **50**, 237-240.
- WEYBREW, B. B. Psychological problems associated with prolonged periods of marine submergence. In N. M. Burns, R. M. Chambers, and E. Hendler (Eds.), *Unusual environments and human behavior*. Glencoe, Ill.: Free Press, 1962.

(Received March 6, 1962)

PERSONALITY CORRELATES OF DELINQUENCY RATE IN A NAVY SAMPLE

ROBERT R. KNAPP¹

United States Naval Personnel Research Activity, San Diego, California

A study to determine whether personality scales measuring social maturity and conformity were related to delinquency rate in a group of 92 Navy brig confinees. The Socialization scale of the California Psychological Inventory and the Conformity scale of the Survey of Interpersonal Values each correlated significantly with a delinquency criterion obtained by correlating the scales against number of offenses committed and partialing out length of service. While the validities obtained for both the Socialization and Conformity scales were in the expected direction, the difference between means for the present delinquent and other nondelinquent groups on the Conformity scale is not in accord with the validity obtained. Differences in mean scores between the present delinquent sample and those presented for high school samples were in the expected direction for the Socialization scale.

Recidivism is a problem of concern in military and civilian settings. In the military, repetition of offenses represents time lost and therefore, the delinquency rate, or frequency with which an individual commits offenses, is of considerable importance in determining whether the individual will be retained in service or administratively separated. Identification of personality characteristics associated with delinquency rate is an important first step in understanding the potentially habitual or frequent offender. Further, the determination of such personality correlates may have implications for remedial or rehabilitative efforts.

A number of previous investigations have been directed toward the identification and measurement of the personality variables associated with delinquency. In a series of investigations, Gough (1954, 1960) and Gough and Peterson (1952) have developed a theoretical formulation of socialized behavior in conjunction with the empirical development of a scale to predict delinquency. In brief, items were selected which maximized differentiation between delinquent and non-

delinquent groups. The resultant personality scale is the Socialization, or delinquency, scale and is incorporated in the California Psychological Inventory (CPI) (Gough, 1957).

In another approach to the identification of predispositional factors in delinquency, previously isolated personality constructs or factors have been found to be related to delinquency. Gordon (1961) noted that mean scores obtained by confined delinquent samples on a scale developed to measure conformity were significantly, and consistently, higher than means for nondelinquent samples. This scale, which is incorporated in the Survey of Interpersonal Values (SIV) (Gordon, 1960), provides a measure of the importance an individual places on such things as "conforming to rules and regulations" and "always doing the approved thing."

Since it was felt that these personality scales of Socialization and Conformity might be particularly relevant to failure to adjust to military discipline, as evidenced by the rate of delinquency, the present investigation was undertaken to determine the relationship between these scales and the delinquency rate obtained from the offense records of a group of military offenders.

METHOD

Subjects

The subjects considered were 100 white, male confinees of the Navy Brig, Marine Corps Barracks,

¹ Now at the United States Navy Medical Neuropsychiatric Research Unit, San Diego, California.

The opinions expressed are those of the author and are not to be construed as being official or in any way representative of the United States Navy.

The author wishes to express his appreciation to Edward F. Alf, Jr., for his valuable advice and assistance throughout the course of the present project.

Naval Station, San Diego. This sample was taken from among individuals entering the brig during a 1-month period. Subjects were selected on the basis of availability of previous offense records. However, incomplete records were found for six of the subjects tested. In addition, two of the subjects had been in the Navy less than 3 months, and thus did not have sufficient service to permit the determination of a meaningful delinquency rate criterion. The remaining 92 subjects constituted the sample under investigation.

The group ranged in age from 17 to 31 years with a mean age of 20.3 years. Educational level was from 7 to 13 years of school completed, with a mean of 9.8 years. Length of service ranged from 7 to 67 months, with a mean of 26.3 months.

Experimental Variables

The SIV and the Socialization scale were administered to groups of about 20 each during the subjects' first week in the brig. The SIV is a forced-choice instrument yielding measures of six values. Validity for only one of these values, Conformity, was hypothesized. Means, standard deviations, and validities of all scales are presented, since the nature of the instrument does not permit administration of the single Conformity scale.

General Classification Test (GCT) scores and educational level (highest grade reached) were also available for all subjects. Since these variables had been found to be associated with delinquency in other studies (e.g., Flyer, 1959) correlations of GCT and education with the delinquency criteria were also obtained.

Delinquency Criteria

Review of the offenses recorded showed that the large majority of offenses committed consisted of unauthorized absence. Other offenses also were primarily military in nature (e.g., disrespect, failure

to obey a lawful order). There were only a very few nonmilitary offenses, such as larceny or assault. In view of the infrequency of offenses of the latter type, an analysis by offense category was considered to be unfeasible.

It was not considered to be adequate to use number of offenses as the criterion, since those individuals who had been in the service longer would have had more opportunity to commit offenses. A simple "number of offenses" criterion might thus be a confounded "length of service" criterion. Consequently, an attempt was made to correct for length of service by devising a "delinquency rate" criterion. The delinquency rate criterion was computed by dividing the individual's number of offenses committed by his years of service. The rate of offenses on an annual basis ranged from .4 to 5.3, with a mean of 1.9 per year.

This delinquency rate criterion, however, was subject to certain methodological objections, the major objection being that in the ratio the denominator variable may be weighted inappropriately (e.g., Asher, 1962). Another method for controlling for length of service would be to partial out this variable from the correlations with number of offenses. These partial correlations against number of offenses might more adequately represent the relations between offense and the predictors with length of service held constant.

Validities in the form of Pearson product-moment correlations against number of offenses with length of service partialled out were computed for the Gough Socialization scale and for the scales of the SIV. Correlations are also presented for these variables against the delinquency rate criterion.

RESULTS

Intercorrelations among personality scales, GCT, educational level, length of service, and number of offenses are presented as a matter

TABLE 1
INTERCORRELATIONS BETWEEN VARIABLES

Variable	2	3	4	5	6	7	8	9	10	11
1. SIV Support	-.06	.32	-.17	-.19	-.35	-.09	.08	-.19	-.03	.15
2. SIV Conformity		-.49	-.40	.48	.10	.20	-.12	-.05	-.01	-.26
3. SIV Recognition			-.30	-.46	.33	.00	.28	.05	.00	.05
4. SIV Independence				-.37	-.17	-.29	-.05	-.09	.03	.17
5. SIV Benevolence					-.44	.21	-.27	.07	-.04	-.18
6. SIV Leadership						.06	.16	.22	.03	.04
7. CPI Socialization							-.01	.18	-.05	-.27
8. GCT								.33	.24	.14
9. Educational level									.33	-.02
10. Length of service										.51
11. Number of offenses										

Note.—*N* = 92.

TABLE 2
MEANS, STANDARD DEVIATIONS,
AND VALIDITIES

Variable	<i>M</i>	<i>SD</i>	Validity ^a	
			Delinquency rate	<i>r</i> _{12.3}
SIV Support	14.14	4.94	.14	.19
SIV Conformity	20.80	6.13	-.24*	-.29**
SIV Recognition	10.67	4.16	.06	.06
SIV Independence	16.52	7.56	.14	.18
SIV Benevolence	16.83	5.37	-.11	-.19
SIV Leadership	10.98	6.12	.00	.02
CPI Socialization	29.34	7.12	-.24*	-.28**
GCT	48.71	7.98	-.07	.02
Educational level	9.84	1.29	-.12	-.23*

Note.—*N* = 92.

^a The *r*_{12.3} column presents correlations of test data and educational level against number of offenses with length of service partialled out. The delinquency rate criterion was determined from the ratio of number of offenses to length of service.

* *p* ≤ .05.

** *p* ≤ .01.

of interest in Table 1. Means, standard deviations, and validities for the predictor variables are presented in Table 2.

Means for the Conformity scale of the SIV were compared with those reported by Gordon (1961) for other delinquent groups. The means of 20.80 obtained for the present sample falls within the range of means reported for state prison, criminally insane, and juvenile delinquent samples, which were 19.1, 22.3, and 19.2, respectively.

The mean score of 29.34 obtained in the present sample on the Socialization scale places this sample between the high school disciplinary problem sample and the county jail inmate sample which had means of 31.25 and 29.27, respectively (Gough, 1960, p. 25).

The mean GCT score and educational level are only slightly below the means generally found for these variables in an unselected Navy population.

Table 2 shows that significant negative relationships exist between the delinquency criteria and the personality scales of Conformity and Socialization. In other words, individuals who score *lower* in social maturity and who place relatively *less* importance on following rules and regulations, as measured by the Socialization and Conformity scales, respectively, are likely to have a higher offense rate during their military service.

Educational level was significantly related, at the .05 confidence level, to number of offenses when length of service is partialled out. Those with lower educational levels tend to have a higher rate of delinquency. While this correlation is low, it does support the finding of Flyer (1959) where those Air Force personnel given unsuitability discharges had a lower mean educational level.

In the present study, aptitude level (as measured by GCT) was not related to the delinquency criteria.

DISCUSSION

The validities obtained for both the Socialization and Conformity scales were in the expected direction, that is, lower Socialization and Conformity scores were associated with higher rates of delinquency.

Further, the Socialization mean obtained for the present sample is lower than the mean reported by Gough (1960) for a male high school sample, as would be expected (29.34 as compared with 36.46). However, the mean obtained for the SIV Conformity scale is higher than the high school mean reported by Gordon (1961) (20.80 as compared with 14.8).

Since the present data were collected under conditions of confinement, this may have contributed to the overall reported greater valuing of conforming behavior. Two possible explanations might be that delinquents in confinement feel that it is important to say they want to conform, or, they may actually feel that it is more important to conform under these conditions than do students in the high school situation. Another explanation of the discrepancy in the Conformity mean may lie in the possible inappropriateness of high school norms for this instrument as applied to military populations.

REFERENCES

- ASHER, W. Statistical problems of the accomplishment quotient. *J. exp. Educ.*, 1962, 30, 285-287.
- FLYER, E. S. Factors relating to discharge for unsuitability among 1956 airmen accessions to the Air Force. *USAF WADC tech. Rep.*, 1959, No. 59-201.
- GORDON, L. V. *Manual for the Survey of Interpersonal Values*. Chicago, Ill.: Science Research Associates, 1960.

- GORDON, L. V. Conformity among the nonconformists. *Psychol. Rep.*, 1961, **8**, 383.
- GOUGH, H. G. Systematic validation of a test for delinquency. *Amer. Psychologist*, 1954, **9**, 381. (Abstract)
- GOUGH, H. G. *Manual for the California Psychological Inventory*. Palo Alto, Calif.: Consulting Psychologists Press, 1957.
- GOUGH, H. G. Theory and measurement of socialization. *J. consult. Psychol.*, 1960, **24**, 23-30.
- GOUGH, H. G., & PETERSON, D. R. The identification and measurement of predispositional factors in crime and delinquency. *J. consult. Psychol.*, 1952, **16**, 207-212.

(Received March 19, 1962)

AN EXPERIMENTAL COMPARISON OF INVENTORY VALIDITY OBTAINED BEFORE AND AFTER WORK EXPERIENCE

WARREN R. GRAHAM AND CECIL D. JOHNSON¹

United States Army Personnel Research Office, Washington, D. C.

Are responses to inventories more valid when obtained before work experience than after work experience? 2 long self-description inventories were administered to 537 soldiers. The same inventories were administered to 372 members of this sample 18 mo. later and 6 mo. after criterion ratings of performance on maneuvers in Germany were obtained. The remaining 155 men could not be retested and were used to cross-validate the results. 19 short personality and interest keys and 2 total score keys were developed using the before-experience responses and then the after-experience responses. 2 of the 19 personality and interest keys and 1 of 2 total score keys showed statistically significant differences between validities for the before- and after-experience responses. The cross-validity of the regression composite based on before-experience responses was .23; on after-experience responses it was .26.

The traditional theory of test validation assumes that testing is done for the purpose of predicting success as evaluated by a specified criterion measure. It is commonly believed that the appropriate procedure requires test keys and/or regression weights that have been developed from tests administered before experience in training or on a job, and prior to the collection of criterion information. Such a research design can be very expensive and time consuming because: (a) two separate data collection operations are necessary (one for the predictor, and another for the criterion), and (b) the time lapse and expense required for validation of follow-up cases usually are large. In addition, attrition between entry testing and criterion measurement usually restricts the range of talent in the sample with consequent reduction of predictive precision.

The difficulties and expenses involved in follow-up validations frequently cause psychologists to seek more economical, less time consuming methods of item analysis and test validation and battery weighting. The design that usually results from these considerations is to administer the tests to a sample that is already trained and experienced so that criterion information can be obtained immediately. This procedure, however, raises

serious questions: (a) Are the resulting validities and weights as predictive as those that could be obtained by follow-up validation of entry testing? (b) Are differences observed between validities obtained before and after experience due to lack of reliability of the variables? (c) Are the validities of variables that are selected and weighted for inclusion in a test composite stable under the influence of training, maturation, and experience? This study seeks answers to the above questions that will be appropriate for applications of psychological inventories.

It is suspected that a soldier entering the Army for the first time undergoes changes in his attitudes, interests, and adjustment behavior as his training and experience increase. It is not known to what degree such changes affect the predictiveness of self-description inventory variables. If it can be shown that equally good predictions can be obtained by validating responses obtained after experience rather than before, then test research can be speeded and can be completed more economically.

METHOD

Samples

The principal samples used for this investigation consisted of 372 infantrymen in the follow-up validation sample, and 155 infantrymen in the hold-out cross-validation sample. The 155 men in the cross-validation sample were soldiers who were tested before experience, but who could not be

¹The opinions expressed in this paper are those of the authors and do not necessarily reflect official Department of the Army policy.

retested after experience for various military reasons. No systematic bias is known to have affected selection of the 155 men of the cross-validation sample. The validation sample of 372 soldiers was tested both before and after experience. Criterion information was based on rated performance on maneuvers in Germany for both samples. The test variables administered to 527 cases from the Tenth Infantry Division at Fort Riley, Kansas, during April 1955. The 372-man validation sample was retested in December 1956. This was about 7 months after the completion of the maneuvers criterion and about 18 months after the testing before experience, at entry into training.

Variables

The predictor instruments were the Interest-Opinion Questionnaire and the Recruit Self-Description Blank which contain a total of 560 items. Each of these instruments produced a total score. Nineteen short keys also were developed by judgment to represent hypothetical personality and interest variables. In addition, the 19 personality and interest keys were weighted by multiple regression to obtain two composite keys, one from before-experience responses, and the other from after-experience responses. The items used in this study were keyed from previous item validation studies, either for combat performance in Korea, or for performance of typical Army jobs, such as mechanic, cook, driver, etc.

The criterion used for the present study was called the Combat Aptitude Score. This measure is a mean rating of combat performance and is made on a seven-point scale by squad members and superiors (cadre). The criterion measures were obtained on performance during combat maneuvers in Germany. The variables are listed in Table 1.

Problem

The primary purpose of this study was to investigate the hypothesis that a battery of inventory variables weighted by regression from before-experience responses will be equally as valid as when weighted from after-experience responses. Four questions are relevant to the hypothesis: (a) Will the validity of responses after experience be significantly greater than zero? (b) Will the validity of responses before experience be significantly greater than zero? (c) Will the validities of each personality and interest key be equal before and after experience? (d) Will the cross-validities be equal for regression-composite keys obtained from testing before and after experience?

Procedure

The general procedure employed for this study was a test-retest design that involved validation of the keys on the 372-case sample before training and again after training and experience in the field. In addition, two multiple regression composites were

obtained from the 372-case sample (one for each administration) and both composites were cross-validated on an independent sample of 155 entering personnel. Results of the experiment were expressed in terms of critical ratios of the differences between the validity coefficients from the 372-case sample, and between the cross-validities of the regression-weighted composites from the 155-case sample. Comparisons between the validity coefficients obtained before and after experience were made by using a statistical test that took into account the correlation between the administrations (Hotelling's t).

In order to evaluate the cross-validities of the regression-weighted composites, it was first determined that the validity of each personality and interest key from the 155-case sample was drawn from the same population of correlations as was its counterpart validity based on the 372-case sample. No significant difference was found between any validity coefficient from the 372-case sample and its counterpart validity from the 155-case sample. The inventories (Variables 21 and 22) were administered about 18 months apart, and there is no apparent reason to suspect that responses made after experience were influenced by responses made before experience. Thus, it seems reasonable to assume that response to an item made during the first administration was experimentally independent of a response to the same item made 18 months later.

Two separate composites of the personality and interest keys were computed from regression weights obtained from each administration of the inventories (before and after experience). These weights were then applied to the hold-out sample of 155 soldiers to estimate the cross-validity of the regression composites. Although one set of regression weights was obtained from testing before experience and the other from testing after experience, the composite scores from both were cross-validated in the entry testing situation (before experience) in order to duplicate operational procedures.

RESULTS

The validity coefficients for each personality or interest key, and for the total scores, are presented in Table 1. It is apparent that a few of the validities changed significantly with experience. Three of the 19 short personality and interest keys showed a significant change, and one of the long inventory keys (Variable 22) had significantly higher validity for responses made after experience. In one case (Variable 1) a significant change occurred for two validities neither of which was significantly different from zero—a trivial result.

The standard partial regression coefficients are also presented in Table 1. It is evident

TABLE 1
COMPARISONS OF RESULTS FROM RESPONSES BEFORE AND AFTER EXPERIENCE

Variable	Number of items	Validity of key			Regression weights		Stability r_{BA}
		Before	After	t_{BA}	Before	After	
Personality and interest keys							
1. Interest in Outdoor Activities	16	-.05	.05	2.12*	-.066	.049	.59
2. Athletics and Outdoors	14	-.08	.00	1.59	-.125	-.169	.49
3. Prudence	13	.19**	.12*	1.20	.107	.067	.35
4. Social Responsibility	10	.15**	.17**	0.33	.054	.102	.34
5. Acceptance of Authority	10	.10**	.16**	0.95	.002	.105	.24
6. Lack of Psychopathic Symptoms	8	.16**	.10*	0.97	.056	-.023	.30
7. Lack of Neurotic Symptoms	13	.11*	.13**	0.36	.008	.039	.38
8. Lack of Hypochondriasis	16	.13**	.09	0.67	.104	-.046	.34
9. Social Initiative	13	.05	.20**	2.89**	.065	.232	.48
10. Social Skills	17	.07	.12*	0.91	-.018	.032	.44
11. Physical Alertness—Activity	14	.01	.07	1.16	-.146	-.025	.50
12. Mental Alertness	12	.09	.10*	0.15	-.011	-.086	.41
13. Rugged Masculinity	19	.10*	.05	0.99	.184	-.001	.52
14. Lack of Anxiety Symptoms	18	.03	.09	0.86	-.107	-.014	.32
15. Mechanical Interests	23	.08	.01	1.44	.082	.009	.56
16. Lack of Aesthetic Interests	13	.14**	.13**	0.23	.123	.076	.51
17. Lack of Office Interests	18	-.01	.19**	3.25**	-.011	.206	.41
18. Lack of Excitability Symptoms	10	.07	.08	0.15	-.026	.020	.42
19. Lack of Avoidance of People	13	.21**	.11*	1.58	.114	.089	.22
Total score keys							
21. Recruit Self-Description Blank	275	.15**	.22**	1.32	—		.43
22. Interest-Opinion Questionnaire	285	.16**	.25**	2.12*	—		.65
Composite keys (Var. 1–19)							
Multiple correlation ^a	270	.36**	.38**	—			—
Cross-validity of composites ^b	270	.23**	.26**	0.78 ^c			.92 ^c

Note.— $N = 372$.
^a $N = 372$.
^b $N = 155$.
^c Between composites for one administration.
* $p \leq .05$.
** $p \leq .01$.

that the influence of experience results in large changes in the magnitudes of some of the regression weights. For example, one of the weights changed from .065 to .232 (Variable 9) which represents a major alteration in the composition of the weighted composite. This suggests that different information is used for predicting from before-experience composites of keyed responses than is used for after-experience composites.

The cross-validities of the regression-weighted composites of all tests must be compared, however, if independent estimates of the validity of the weighted composites are to be determined. The cross-validity of the composite based on before-experience responses was $r = .23$. The cross-validity of the weighted composite based on after-experience responses was .26. Thus, weights from after-experience responses produced as much cross-

validity as did weights obtained from the customary before-experience responses (follow-up design). The difference between the cross-validities of the composites is not significant ($t_{BA} = .78$, $r_{BA} = .92$).

It is concluded that the use of after-experience responses can lead to erroneous conclusions concerning the validity of single, independently evaluated variables. The risk might be considerably reduced, however, if reasonably large numbers of variables were combined to produce a regression-weighted composite score. The use of such a composite in this experiment resulted in compensatory weighting of information in such a way as to permit the same set of variables to predict equally well from either before-experience or after-experience responses.

(Received March 29, 1962)

EFFECTS OF AUTHORITARIANISM ON VIGILANCE PERFORMANCE¹

BRUCE O. BERGUM AND DONALD J. LEHR

United States Army Air Defense Human Research Unit, Fort Bliss, Texas

An experiment was performed on the effects of authoritarian monitoring conditions upon vigilance performance. Two groups of 20 Ss each were employed. One group worked at a light monitoring task for a period of 135 min. without rest and alone. The second group worked at the same task for the same amount of time but was observed by either a commissioned or noncommissioned officer according to a random visiting schedule. Signal rate was 12 signals per hr. The results indicated a highly significant facilitation of detection performance resulting from observation by the officers. It was suggested that these conditions represent an extreme point along a dimension of perceived threat to the monitor.

In a study on the effects of paired-monitoring Bergum and Lehr (1962b) demonstrated a slight, but statistically nonsignificant, tendency for the presence of a second operator to facilitate individual vigilance performance. In this case, the second individual was a peer. When the second individual represents some form of higher status, however, vigilance performance tends to be significantly improved. Thus, Fraser (1953) found that when the experimenter remained in the testing room the operators performed significantly better than when he was not present. These results suggest an underlying continuum relating to the amount of threat represented by the presence of a second individual. The present study was designed to test a more extreme point along this hypothesized continuum by subjecting enlisted military personnel to observation by superior ranking personnel. It was hypothesized that these conditions would result in a relatively greater improvement in performance than that demonstrated by Fraser.

PROCEDURE

Subjects. Forty National Guard trainees from the Army Training Center, Fort Bliss, served as the

¹ The research reported in this study was performed for the Human Resources Research Office of the George Washington University under contract to the Department of the Army. The opinions expressed in this report are solely those of the authors and do not necessarily reflect those of the sponsoring agency.

The authors wish to express their appreciation to David Cooper and John Mullins for their generous assistance in the conduct of this experiment.

subjects (Ss) in this experiment. These Ss were between the ages of 18 and 26 years and had normal (20/20) vision.

Apparatus. The apparatus employed in this study was identical to that employed in an earlier study (Bergum & Lehr, 1962a) and included four test booths, a circular light panel, and an intercom network. Single response pushbuttons were located in each booth and both signals and responses were automatically recorded on paper tape. The signal rate was 12 signals per hour.

Conditions. Two conditions, permissive and authoritarian, were employed. The Ss were randomly assigned to the two conditions. All Ss in both groups followed the same general procedure. This included a 20-minute pretest session, followed by a 10-minute rest period, followed by 5 contiguous 27-minute periods for a total of 135 minutes of continuous testing in the booths. All Ss were required to deposit their personal timepieces with the experimenters before the experiment began.

The 20 Ss in the permissive condition were instructed to make themselves comfortable in the booths. They were told that for purposes of the research it was important that they detect as many signals as possible, but that they could be free to do anything they desired that would not interfere with this process.

The 20 Ss in the authoritarian group were informed that from time to time a Lieutenant Colonel (or Master Sergeant) would visit them in the booths to observe their performance. The Colonel and the Sergeant each visited 10 Ss during the course of the testing. These observers entered the booths only during those intervals when signals were not programmed, and remained in the booths until at least one signal had occurred. Failures to detect signals were pointed out to the Ss by the observers and conversations were generally held to a minimum. Visits were made according to a prearranged, but apparently random, schedule in which frequency and intervals between visits were counterbalanced across time periods. Each S was visited approximately four times in the course of the testing.

TABLE 1
MEAN PERCENTAGES OF CORRECT DETECTIONS

Condition	Time period					Average
	1	2	3	4	5	
Authoritarian	97%	77%	72%	71%•	80%	79%
Permissive	77%	50%	30%	34%	34%	45%
Average	87%	63%	51%	53%	57%	

RESULTS

The results are presented separately in terms of pretest performance and main variable effects.

Pretest Performance. A Kruskal-Wallis analysis of variance of pretest performance yielded a nonsignificant H of 1.15 between the groups. Similarly, comparisons of the groups in terms of age and amount of schooling yielded nonsignificant H s of .0002 and 1.47, respectively. It was concluded that no relevant differences existed between the groups prior to experimental treatment.

Authoritarianism. The mean percentages of correct detections for each group at each of the five time periods are presented in Table 1. Both groups demonstrated decreases in performance between the first and second periods, with the permissive group showing an additional decrease in the third period. Performance tended to level off by the end of testing

in both groups, but the authoritarian group tended to be superior throughout all time periods.

An arc-sine transformation was made on the data and the transformed scores subjected to a Trials \times Conditions analysis of variance. The results of this analysis are presented in Table 2. The comparison of conditions yielded an F of 13.80, $p < .005$ for 1 and 28 df . Comparison of the time periods yielded an F of 17.34, $p < .005$ for 4 and 112 df . The interaction term was not significant. A separate analysis of the authoritarian data in terms of the effects of the individual military observers yielded a nonsignificant F of less than unity.

These results indicate that vigilance performance is significantly superior under authoritarian monitoring conditions (as presently defined), but that this effect is not sufficiently powerful to eliminate a significant performance decrement over time.

TABLE 2
ANALYSIS OF VARIANCE FOR
TRANSFORMED DATA

Source	SS	df	MS	F
Between Conditions	31,059	1	31,059	13.80**
Between Ss	63,030	28	2,251	
Total	94,089	29		
Between Periods	18,739	4	4,685	17.34**
P \times C	2,069	4	517	1.91*
Ss \times P	30,265	112	270	
Total	51,073	120		
Total	145,162	149		

Note.—P = Periods, C = Conditions.

* $p > .05$.

** $p < .005$.

DISCUSSION

The present results confirm the expectation that detection performance can be maintained at relatively high levels under authoritarian conditions. Of particular interest is the magnitude of the difference in performance between the two groups. This difference (46% in the final period of testing) is even greater than that demonstrated by the authors (Bergum & Lehr, 1962a) in an earlier study of interpolated rest periods in which the rest group maintained near-perfect performance throughout the testing period. While the present authoritarian group did demonstrate a performance decrement, the very poor performance of the control group suggests that had the "easier" task employed in the earlier

study been employed in the present study the authoritarian condition might have yielded similar near-perfect performance.

In terms of the hypothesis that the present conditions represent an extreme point along a continuum relating to the amount of threat represented by the presence of a second individual, comparison of the present study with Fraser's study (1953) and the Bergum and Lehr pairing study (1962b) in terms of the ratio of errors between control and experimental groups for comparable 1-hour periods indicates an ordering in the predicted direction. These ratios are: pairing, .689; experimenter present, .410; and ranking military observer, .363. As anticipated, Fraser's results

lie slightly closer to the authoritarian than to the pairing condition results and this relationship is in conformity with intuitive evaluations of the conditions in terms of the amount of perceived threat in these situations.

REFERENCES

- BERGUM, B., & LEHR, D. Vigilance performance as a function of interpolated rest. *J. appl. Psychol.*, 1962, **46**, 425-427. (a)
- BERGUM, B., & LEHR, D. Vigilance performance as a function of paired monitoring. *J. appl. Psychol.*, 1962, **46**, 341-343. (b)
- FRASER, D. C. The relation of an environmental variable to performance in a prolonged visual task. *Quart. J. exp. Psychol.*, 1953, **5**, 31-32.

(Received May 7, 1962)

HUMAN RELATIONS KNOWLEDGE AND SOCIAL DISTANCE SET IN SUPERVISORS

FRANCES M. CARP

Trinity University

BART M. VITOLA¹

San Antonio Post Office, Texas

AND FRANK L. McLANATHAN

Headquarters, United States Air Force, Washington, D. C.

Postal supervisors' knowledge of human relations as measured by an adapted Air Force test and their perceptual set for social distance as measured by Fiedler's ASo Scale correlated significantly with productivity of subordinates as measured by the United States Postal Department Work Production Standard averaged over the period of 1 year. Effective leaders knew proper human relations practices and their perceptual set enabled them to maintain optimal psychological distance from subordinates, neither so close that they were hampered by emotional ties nor so distant that they lost emotional contact. Results suggest that selection of Postal Department supervisors would be improved by eliminating candidates whose scores fall beyond 1 sigma from the mean on the social distance scale; and from those remaining, taking the number required with the highest human relations knowledge scores.

Knowledge of good human relations practices should benefit a supervisor. However a person may give "correct answers" inconsistent with behavior. It is important to clarify the relationship between knowledge as measured on verbal tests and effectiveness in supervision.

Studies of Fiedler and others (Cleven & Fiedler, 1956; Cronbach, 1955; Fiedler, 1955, 1958) suggest that performance of the supervisory role depends also upon ability to maintain optimal psychological distance from subordinates. It must not be so great that the leader lacks contact with other members nor so small that he is unable to deal objectively with their performance.

This study examines the potential usefulness of one human relations knowledge test and one measure of perceptual set for social distance by determining their relationship with an independently established criterion of supervisory effectiveness.

METHOD

Subjects. Forty-one of 60 supervisors in the San Antonio, Texas, Post Office² participated. Of the

¹ Now with the Air Force Personnel Laboratory.

² The authors are most grateful to Post Office personnel for access to subjects and criterion.

remainder, 4 could not be contacted, 10 expressed disapproval of the study, and 5 accepted the materials but did not return them. Participants had been on supervisory status from 1 to 25 years, non-participants from 10 to 20 years.

Criterion. The United States Postal Department routinely records a Productivity Index (PI) for each unit of workers which compares the volume of mail it processes to a Work Production Standard established in a motion-time study. The criterion for the present study was the daily PI of each supervisor's unit averaged over the period of 1 year. Scores ranged from 60 (i.e., 60% of the Standard) to 90 with a mean of 71.38 and a standard deviation of 7.74.

Human relations knowledge test. An Air Force test (SKT)³ comprised of situational problems with multiple-choice answers was reduced from 150 to 117 items pertaining to general supervision. Items requiring military knowledge were deleted and Postal Department terminology substituted for that of the Air Force. Sample items:

51. The relationship between a supervisor and his subordinates should be based primarily upon
 - (a) respect for the greater experience of the supervisor
 - (b) respect for the superior technical knowledge of the supervisor
 - (c) sharing responsibility in the unit
 - (d) confidence in the supervisor and his decisions

³ We also thank the Air Force Personnel Laboratory for permission to use a test as yet unpublished.

63. Which one of the following is the best reason for the proper use of controls in an organization?
 - (a) to maintain discipline in the organization
 - (b) to eliminate poor workers and retain good ones
 - (c) to assure quality and quantity of production
 - (d) to regulate the work flow in a unit
67. A supervisor desires to keep his personnel informed of the current activities of other sections that are performing work similar to that of his own. How can he best pass this information to his subordinates?
 - (a) by scheduling periodic staff conferences
 - (b) by circulating brief written reports
 - (c) by asking his most able subordinates to convey this information to other subordinates
 - (d) by posting this information on the bulletin board of his section
75. A tour superintendent is in charge of a large group of personnel who work in widely separated localities. If the tour superintendent has difficulty in controlling the work, which of the following actions should he take?
 - (a) confer frequently with subordinates
 - (b) maintain strict control
 - (c) adjust work schedules
 - (d) delegate authority
89. Which one of the following is the first factor to consider in the preparation of a training problem?
 - (a) equipment available
 - (b) instructors available
 - (c) scheduling the training
 - (d) objective of the mission
103. The supervisor who is rating a subordinate for efficiency should be careful to make sure that his rating is affected LEAST by his subordinate's
 - (a) reaction to criticism
 - (b) ability to cooperate with others
 - (c) reaction to emergency situations
 - (d) popularity among his fellow workers
109. When an employee fails to carry out an order, the first possibility that the supervisor should consider is that the employee
 - (a) deliberately disobeyed
 - (b) did not understand
 - (c) forgot to carry it out
 - (d) was unable to carry it out

Each item in the Air Force form had a validity coefficient significant at the .01 level against academic grades in noncommissioned officer academies and supervisors' ratings. Raw scores for the postal group ranged from 50 to 90 with a mean of 77.84 and a standard deviation of 10.68.

Psychological distance measure. Fiedler's assumed similarity of opposites (ASo) as measured by the D score (Cronbach & Glaser, 1951) was used as a measure of the perceptual set which largely influences the social distance established between leader and subordinate. ASo test questions are phrased in terms of perception of the "least preferred" co-worker. High ASo scores show dependence on others, concern with the feelings of others, and unwillingness to reject a person with whom one cannot accomplish a task. Low ASo scores indicate independence of others, lack of concern for the feelings of others, and readiness to reject others considered to be poor co-workers. Most effective supervisors should score in the middle range, being psychologically close enough to group members to maintain contact and far enough to allow objective reaction to poor performance. Scores for these subjects ranged from 10.2 to 39.3 with a mean of 22.94 and a standard deviation of 8.04.

Hypotheses.

1. There is a significant relationship between SKT scores of supervisors and PIs of groups working under them.
2. Optimal ASo scores of supervisors lie in the middle range; work groups whose supervisors have ASo scores within 1 sigma of the mean have higher PIs than those whose leaders have ASo scores beyond those limits.
3. Knowledge of good supervisory practices and set for social distance, used together, are more closely related to success than either of the variables alone.
 - (a) The multiple correlation of SKT and ASo scores with the criterion is higher than either of the single correlations.
 - (b) Independence of contributions of SKT and ASo scores is shown by a low and statistically insignificant correlation between them.
 - (c) Perhaps most practically, after elimination on the basis of one measure, further improvement in selection is possible, using the second. After all subjects with ASo scores falling beyond 1 sigma from the mean are eliminated, the remainder show a significant correlation between SKT scores and the criterion.

RESULTS

First Hypothesis. Supervisors' SKT scores were related to PIs of groups working under them ($r = .65, p < .01$).

Second Hypothesis. Groups working under supervisors who scored within 1 sigma of the mean on ASo were more productive than groups working under supervisors who scored beyond these limits (point biserial $r = .70, p < .01$). The possibility of false negatives and false positives seems negligible: a PI of 68 was the highest obtained by any subject scoring beyond 1 sigma of the mean on ASo and it was the lowest PI for any subject within that middle range.

Third Hypothesis. (a) Multiple correlation of SKT and ASo scores with PIs was higher than either of the single correlations ($R_{1.23} = .84$; $R-r = .14$; $.05 < p < .10$). (b) The correlation between SKT and ASo scores for all subjects was low and statistically insignificant ($r = .11$). (c) Subjects whose ASo scores fell beyond 1 sigma of the mean were eliminated. For the remaining 23 subjects, correlation between ASo and human relations knowledge scores was similar to that for the entire group ($r = .64$, $p < .01$). After elimination on the basis of ASo scores, SKT scores still showed meaningful relationship with productivity and so should improve selection of supervisors. The two measures made relatively independent and complementary contributions to leadership effectiveness.

DISCUSSION

Supervisors of effective postal teams scored high in knowledge of human relations principles and perceived others in such a way that they could establish optimal psychological distance between themselves and their subordinates. Scores on a test of perceptual set for least preferred co-workers suggest that effective supervisors stay psychologically close enough to maintain contact with group members and far enough to deal objectively with poor performance. These findings support Fiedler's contention that the leader should neither perceive himself as so distant from the group that it receives no support from him, nor so close to it that he is hampered by emotional ties.

Regarding the advantage of using both

techniques in supervisor selection, (a) the multiple correlation exceeded both single correlations with the criterion measure, (b) relative independence of ASo and SKT scores was demonstrated, and (c) the utility of eliminating on the basis of extreme ASo scores and then selecting by SKT scores was clear-cut. As an example of the last point, supervisors in this study had been selected from the clerk-carrier force, a few at a time as openings occurred, almost exclusively on the basis of seniority. The mean PI for all 41 was 71. For the 23 whose scores fell within 1 sigma of the mean on ASo, the PI was 77.37. The 4 men of this 23 who scored highest on SKT had an average PI of 84. Since there is normally an excess of candidates for promotion to supervisory levels, the advantage of selecting in this way, rather than on the basis of seniority, seems obvious and would be administratively simple.

REFERENCES

- CLEVEN, W. A., & FIEDLER, F. E. Interpersonal perceptions of open-hearth foremen and steel productions. *J. appl. Psychol.*, 1956, **40**, 312-314.
- CRONBACH, L. J. Processes affecting scores on understanding of others and "assumed similarity." *Psychol. Bull.*, 1955, **52**, 177-193.
- CRONBACH, L. J., & GLASER, GOLDINE C. Similarity between persons and related problems of profile analysis. Technical Report No. 2, 1951, University of Illinois, Project N6-ori-07B5, Office of Naval Research. (Mimeo)
- FIEDLER, F. E. The influence of leader-laymen relations on combat crew effectiveness. *J. abnorm. soc. Psychol.*, 1955, **51**, 227-235.
- FIEDLER, F. E. Leader attitudes and group effectiveness. Report, 1958, University of Illinois.

(Received May 14, 1962)

Journal of Applied Psychology

KENNETH E. CLARK, Editor
University of Colorado

Table of Contents

Moderating Effects and Differential Reliability and Validity: Edwin E. Ghiselli.....	81
Development of a Forced-Choice Rating Scale for Engineer Evaluation: J. Richard Lepkowski.....	87
Factors Affecting Employee Selection in Two Cultures: Harry C. Triandis.....	89
Personal Adjustment and Academic Predictability among College Freshmen: L. Bryce Andersen and Patricia A. Spencer.....	97
Prediction of Attendant Tenure: Earl C. Butterfield and Sue A. Warren.....	101
Team Creativity as a Function of the Creativity of the Members: Harry C. Triandis, Alan R. Bass, Robert B. Ewen, and Eleanor Hall Mikesell.....	104
Reworded versus New Interest Items: Edward K. Strong, Jr.....	111
Influence of Digit Grouping on Memory for Telephone Numbers: Francis T. Severin and Marilyn K. Rigby.....	117
Effect of Rehearsal of Temporal and Spatial Aspects on the Long-Term Retention of a Procedural Skill: James C. Naylor and George E. Briggs.....	120
Chance on SVIB: Dice or Men?: David Campbell.....	127
Vigilance Performance under Conditions of Redundant and Nonredundant Signal Presentation: William C. Osborn, Richard W. Sheldon, and Robert A. Baker.....	130
Self-Esteem and the Diffusion of Leadership Style: David G. Bowers.....	135
Job Attitudes in Management: II. Perceived Importance of Needs as a Function of Job Level: Lyman W. Porter.....	141
Retranslation of Expectations: An Approach to the Construction of Unambiguous Anchors for Rating Scales: Patricia Cain Smith and L. M. Kendall.....	149
Self-Confidence as a Response Set: Cecil J. Mullins.....	156
Two Approaches to the Prediction of Group Responses: Rossall J. Johnson.....	158

American Psychological Association

Consulting Editors

GEORGE E. BRIGGS, *Ohio State University*
NORMAN FREDERIKSEN, *Educational Testing Service*
LEONARD D. GOODSTEIN, *University of Iowa*
EDWIN R. HENRY, *Standard Oil Company of New Jersey*
JOHN HOLLAND, *National Merit Scholarship Corporation*
CLIFFORD E. JURGENSEN, *Minneapolis Gas Company*

LAURENCE S. MCGAUGHRAN, *University of Houston*
QUINN MCNEMAR, *Stanford University*
HAROLD F. ROTHE, *Beloit Corporation*
THOMAS A. RYAN, *Cornell University*
ALEXANDER G. WESMAN, *Psychological Corporation*
CLARK L. WILSON, *Batten, Barton, Durstine, and Osborn, Incorporated*

This journal gives primary consideration to original investigations in any field of applied psychology except clinical psychology, although a descriptive or theoretical article may be accepted if it represents a special contribution in an applied field. Quantitative investigations of interest or value to psychologists working in the following broad fields will be considered: vocational and educational prognosis, diagnosis, and guidance at the secondary and college level; personnel research in business, industry, and government; engineering psychology; industrial working conditions; research on opinion and morale factors; job analysis and classification research; market and advertising research.

Manuscripts must be accompanied by an abstract of 100–120 words typed on a separate sheet of paper. The abstract should conform to the style of *Psychological Abstracts*. Detailed instructions for preparation of the abstracts appeared in the *American Psychologist* (1961, 16, 833), or they may be obtained from the Editor or from the APA Central Office.

Manuscripts should be addressed to the Editor:

Dr. Kenneth E. Clark
Office of the Dean
College of Arts and Sciences
University of Colorado
Boulder, Colorado

All manuscripts must be submitted in duplicate. Original figures are prepared for publication; duplicate figures may be photographs or pencil-drawn copies.

Manuscripts must conform to the style requirements described in the *Publication Manual of the American Psychological Association*.

The following policies govern reprinting of materials copyrighted in APA journals: (a) to require approval to reprint for tables and figures, and for text only if more than 500 words [total from one article] in length; (b) to continue approval to reprint, contingent on the author's approval, for articles reprinted in whole or in major part; (c) to negotiate where possible the dedication of royalties from commercial publishers to the American Psychological Foundation.

Journal of Applied Psychology

Published bimonthly by the
American Psychological Association
Prince and Lemon Sts., Lancaster, Pa.
and 1333 Sixteenth Street N. W.
Washington 6, D. C.

\$10.00 per volume

\$2.00 per issue

HELEN ORR
Managing Editor

ELIZABETH S. REED
Advertising Manager

VIRGINIA RICHARDS
Editorial Assistant

Subscriptions, orders, and business communications should be addressed to the American Psychological Association, 1333 Sixteenth St. N.W., Washington 6, D. C. Address changes must reach the subscription office by the tenth of the month to take effect the following month. Undelivered copies resulting from address changes will not be replaced; subscribers should notify the post office that they will guarantee second-class forwarding postage. Other claims for undelivered copies must be made within four months of publication.

Second class postage paid at Lancaster, Pennsylvania and at additional mailing places.

© 1963 by the American Psychological Association, Inc.

MODERATING EFFECTS AND DIFFERENTIAL RELIABILITY AND VALIDITY

EDWIN E. GHISELLI

University of California

Classic psychometric theory holds that errors of measurement and of prediction are of the same magnitude for all individuals. Interactive effects are not recognized, and the psychological structure of all individuals is taken to be the same. To increase reliability and validity of measurement, then, attention is entirely focused on improvement of measuring devices. However, a substantial body of evidence indicates there are systematic individual differences in error, and in the importance a given trait has in determining a particular performance. Reliability and validity of measurement can be increased by the use of moderator variables which predict individual differences in error and in the importance of traits.

For more than half a century the notions of Yule and Spearman have dominated theoretical formulations in psychometrics. Pursuant to these classical notions errors are taken to be random and scores are combined additively. The possibility of interactive effects among variables is not recognized. Because in the linear combination of variables their weights are the same for all individuals, it is presumed that the psychological structure of all individuals is precisely the same.

On any one administration of a test, error scores are taken to vary from individual to individual. Hence for some individuals the error of measurement is smaller and for others it is larger. However, over many parallel tests the standard deviation of the errors is taken to be the same for all individuals. More correctly it should be said that as the number of parallel tests increases without limit the standard error of measurement approaches the same value for all individuals. Hence it is concluded that for a given test all individuals are measured with the same degree of reliability.

Similarly, for any one administration of a given criterion and a test the error with which the test predicts the criterion is taken to vary from individual to individual. Hence for

some individuals the error of prediction is smaller and for others it is larger. However, over many parallel criteria and tests the standard deviation of the errors is taken to be the same for all individuals. Again, more correctly it should be said that as the number of parallel criteria and tests increases without limit the standard error of prediction approaches the same value for all individuals. Hence it is concluded that for a given criterion and test all individuals are measured with the same degree of validity.

Because it is held that errors are random and equal for all individuals, and scores are additive with no interactive effects, it follows that neither reliability nor validity can be improved by selecting out from the total group those individuals for whom error is smaller. Reliability can be improved only by increasing the number of measurements, elimination of elements of lesser reliability, or better "house-keeping" procedures designed to reduce random error. Validity can be improved only by increasing the reliability of the criterion and predictor, or adding other predictors which cover aspects of the criterion not measured by the original predictor or aspects of the original predictor which are independent of the criterion.

Classic psychometric theory deals with a large number of sets of measurements, but let us concern ourselves only with two as we ordinarily do in the practical situation. Consider the bivariate distribution of scores on two variables where the relationship is less than unity. The two variables can be either two parallel tests or a criterion and a predictor. Running from the upper right hand to the lower left hand of the bivariate distribution chart is a group of individuals for whom scores are highly related. For this group the differences, regardless of sign, between standard scores on the two variables and the error of measurement or of prediction are small. For the remainder of the individuals the differences between the two standard scores are greater and hence the error is greater. If it could be demonstrated that these differences, or some other measure of error such as the standard error or the correlation coefficient, were related to another variable then some modification of classic psychometric theory would appear to be in order. Ghiselli (1960b) has called this other variable a predictability variable, but Saunders (1956) has better termed it a moderator thus drawing attention to the interactive effects.

Fisk and Rice (1955) have summarized early evidence indicating that individual error of measurement may be predicted by a moderator. More recent demonstrations are provided by Fisk (1957a) and Berdie (1961). Stagner (1933), Abelson (1952), Hoyt and Norman (1954), Holzman, Brown, and Farquhar (1954), Fredericksen and Melville (1954), Saunders (1956), Ghiselli (1956), Fredericksen and Gilbert (1960), and Ghiselli (1960b), among others, have shown that the error of prediction itself may be predicted by a moderator.

Using the procedure he employed to study moderating effects on validity (Ghiselli, 1960b), the present author further examined moderating effects in reliability of measurement. Two parallel forms of a complex reactions test were administered to 775 semi-skilled workers, 517 of whom were used as an experimental group and 258 as a cross-validation group. Each person took both forms of the test on the same occasion. For each member of the experimental group the differ-

ence, regardless of sign, between standard scores on the two forms was determined. It was found that age, education, and scores on a tapping and dotting test were related to these differences. A combination of scores on these variables was taken to form a moderator. The reliability coefficient, the correlation between the two parallel forms, was .92 for the entire group, whereas for the 9% of subjects earning the lowest moderator scores it was only .82 and for the 15% earning the highest moderator scores it was .97.

Three other instances of moderating effects in validity also may be described. A 64-item forced-choice inventory was administered to 96 factory workers on whom criterion scores in the form of supervisors' ratings were available. Seventeen of the items were used in a scale designed to measure "Sociometric Popularity." Half of the workers were used as an experimental group and half as a cross-validation group. For the experimental group the differences, regardless of sign, were determined for each individual between standard criterion and standard test scores. For the 47 items not used in the predictor scale an item analysis was performed against these differences. Responses to 15 of these items were found to be significantly related to the differences and were formed into a moderator to be applied to the cross-validation group. For the entire cross-validation group the validity of the scale as given by the Pearsonian coefficient was $-.01$, whereas for the 19% earning the lowest moderator scores the validity was $-.47$ and for the 32% earning the highest scores it was $.39$.

A group of 144 foremen and a group of 154 executives were rated by their superiors who divided them into two groups, the more and the less successful. Forty percent of the foremen and 42% of the executives were placed in the upper category. All men took the 64-item forced-choice inventory, 24 items of which were scored in a supervisory ability scale. Critical test scores had been set for the two groups such that the scores of 43% of the foremen fell above their critical score and 39% of the executives fell above theirs. In both groups half of the men were placed in an experimental group and the other half in a cross-validation group. Each of the two experi-

mental groups were further subdivided into two groups, one consisting of those individuals who were either high or low on both variables, the "on quadrant" cases, and those who were high on one variable and low on the other, the "off quadrant" cases. An item analysis was performed using the 40 nonscored items. For the foremen, 9 items, and for the executives, 13 items, were found to differentiate significantly between the on quadrant and off quadrant cases, and they were formed into two moderators to be applied to the cross-validation groups. Only three items were common to both moderators. For the entire cross-validation group of foremen the phi coefficient between the criterion and predictor was .26, whereas for the 17% earning the lowest moderator scores the coefficient was .10 and for the highest it was .41. For the entire group of executives the phi coefficient between the criterion and predictor was .41, whereas for the 21% earning the lowest moderator scores the coefficient was .10 and for the 26% earning the highest scores it was .68.

There is, then, a substantial body of evidence indicating that it is possible to predict individual error of measurement and error of prediction. Clearly those individuals for whom a test has a greater degree of reliability or validity can be systematically differentiated from those for whom it has a lesser degree. The higher the cutting score on the moderator is set the higher is the reliability or validity of the test for those individuals who fall above it. The choice of a cutting score is a matter of how many individuals one is willing to eliminate in order to achieve a higher degree of reliability or validity.

Even recognizing that it is possible to differentiate within a group those individuals whose scores are more reliable from those whose scores are less reliable, the practical value of such a differentiation might well be questioned. However, purely for descriptive purposes it might be desirable to know how reliably an individual is measured. Thus if administrative decisions are to be made on the basis of some test or other measuring device, it would be very helpful in borderline cases to have some indication of whether a given individual is measured with a small or large error. Furthermore, with a lower error

of measurement validity should be enhanced. Classic psychometric theory itself teaches this, and Berdie (1961) has given an empirical demonstration. Finally, in some situations it might be highly desirable to be able to predict the extent of intraindividual variability in performance. In personnel selection ordinarily the aim is to pick out those individuals whose performance is high. But for planning purposes or to insure the smooth flow of work it might be equally important to select individuals whose rate of work does not vary greatly from one period to another, that is, has a high degree of self-consistency or reliability.

The case for validity is much clearer since a reduction in error means more accurate prediction and hence the selection of higher performing individuals. But even here the use of moderators might be criticized on the grounds that it necessitates the elimination of a substantial proportion of cases from the appraisal procedures which in turn eliminates even more. However, this is not necessarily the case. Ghiselli (1956) has shown that if a given percentage of individuals is to be selected and the rest eliminated, selecting that given percentage on the combined basis of their moderator and test scores yields a substantially superior group of individuals than that selected on the basis of test scores alone.

Furthermore, in some instances, especially those where the validity for the total group is low or zero, for those individuals who earn low scores on the moderator and who might therefore be eliminated from the appraisal, the validity coefficient may be of respectable magnitude and negative. For example, with the factory workers mentioned earlier the validity of the predictor for the 32% earning the highest moderator scores was .39. But in addition, for the 19% who earned the lowest moderator scores the validity coefficient was $-.47$. So for these latter individuals high predictor scores were associated with low criterion scores. Consequently for half of the total group, the 32% earning high moderator scores and 19% earning low moderator scores, the validity of the predictor is of the order of .40. It may seem peculiar, but a given score on a test may indicate the promise of success for some individuals, whereas for others it may indicate the likelihood of failure.

Another way to use a moderator, and a way which permits an assessment of all individuals, is to determine which of two predictors to use in selection. Ghiselli (1960a) accomplished this by determining for each individual the difference between his standard criterion score on Predictor 1, and also the difference between his standard criterion score and his standard score on Predictor 2. These differences were taken regardless of sign. Thus for each individual the difference between the two differences was determined. For a given individual a positive difference indicates that the one test gives the better prediction and a negative difference that the other is better. Moderators were then developed which were related to these differences and could be applied to cross-validation groups. For some individuals the moderator selects Predictor 1 and for the others it selects Predictor 2, but the standard scores for all individuals are thrown together regardless of predictor.

Ghiselli presents three instances where this proved to be an effective procedure. In one the validity coefficients for two predictors for a particular criterion were .02 and .20, and using predictors selected by the moderator the coefficient was .33. In another instance the validity coefficients of the two predictors were .55 and .61, with predictors selected by the moderator having a validity of .73. Finally, in an instance where the two validity coefficients were .17 and .51, using a moderator to select the better predictor for each individual gave a coefficient of .73.

Obviously the nature of the traits which function as moderators is a matter of considerable importance. Clearly it would be most helpful if all moderators had characteristics in common. Some of the research does suggest that "undesirable" traits such as a lack of personality integration and low motivation are associated with larger error. But certainly many of the traits which have been found to be effective as moderators are of quite a different sort such as age, education, type of interest, and manual dexterity. In a number of Ghiselli's studies moderators were developed through item analyses of the same inventory so that similarity of items which form different moderators can be examined. His results indicate that there is a high degree

of specificity. With two different tests predicting the same criterion for a given group, and with the same test predicting different though similar criteria for two different groups, the items which form the moderators are quite different. While moderator variables are by no means as elusive as suppressor variables, since so many investigators have been able to find or develop them, they do seem to be just as specific. It would therefore appear to be impossible to state any general principles about the nature of the traits which act as moderators. Of course when the bivariate distribution of criterion and test scores is heteroscedastic then test scores themselves serve as a moderator because they are related to error of prediction (Kahneman & Ghiselli, 1962).

Some of the findings indicate that the relationships between moderator scores and scores both on criteria and predictors are quite low. Therefore, they do not add to prediction in a multiple correlation sense. The contribution of moderators is of an entirely different order, differentiating those individuals for whom error is smaller from those for whom it is larger. Their contribution, then, is unique.

As has been seen, there is a substantial body of empirical evidence indicating that moderator effects do occur. Convincing though these findings may be, one would be much more persuaded of moderating effects if some theoretical foundation of them were provided. It could be, as Saunders (1956) and Berdie (1961) have suggested, that moderators operate by sorting a heterogeneous aggregation of individuals into homogeneous groups. The magnitude and pattern of intercorrelations among variables, and hence reliability and validity, vary from group to group. Heterogeneity would be indicated by systematic variation of error from individual to individual whereas homogeneity would be indicated by all individuals having the same error. This notion permits retention of the classic psychometric concepts of randomness of errors and the linear combination of variables. What it adds is the admission that the magnitude of error and the differential weights carried by the components in a composite, the psychological structure, may vary from group to group. However, within a group the error of measurement and of prediction, and the relative

weights carried by a set of tests in predicting a criterion are the same for all individuals.

Thus, women might be less distracted than men by environmental changes during a testing session and hence be more reliably measured. In this case, sex would moderate error of measurement. Intelligence might be more related to grades in engineering school for those students who have substantial interest in engineering than for those whose interest is low. Engineering interest, then, moderates error of prediction. For younger factory workers finger dexterity might be more important than spatial ability in predicting rate of production on the job, and the reverse might be true for older workers. So age would function to moderate the relative weights finger dexterity and spatial ability have in predicting rate of production.

This notion that moderators sort heterogeneous aggregations of individuals into homogeneous groups is a very useful way of conceptualizing moderator effects. It focuses attention on the kinds of differences which exist among individuals who in some given respect are homogeneous thereby suggesting types of moderators. Furthermore, it does little violence to classic psychometric theory. However, it presumes individuals can be divided into clear and distinct classes. Yet in actual practice moderators distribute individuals along a continuum. Individuals are not sorted into separate classes and a "group" is merely those individuals who fall at the same point on the continuum.

Another possible explanation of moderator effects is that the common elements which account for the correlation between two variables differ from individual to individual rather than just from group to group. What in the first point of view were considered as classes are now thought of as class intervals. Interactions among variables of the sort proposed by Lee (1961) are involved.

Error of measurement would be taken as varying from being quite small for some individuals to being quite large for others. Consequently error scores would carry less weight in determining fallible scores for some individuals than for others. Obviously a necessary condition is that individual differences in error scores possess some consistency or reliability

over parallel tests. Evidence supporting this is provided by Fisk and Rice (1955), Fisk (1957b), and Berdie (1961). Such a position would not require that all variation commonly termed error of measurement is predictable by the moderator, but only a portion of it. The remainder would still be thought of as being random error. The reliability coefficient, then, would be an average description of precision of measurement.

The importance of a given trait in determining performance on some criterion is taken to differ among individuals. The trait varies from being of prime importance in determining criterion performance for some individuals to being of little or no importance for others. At the one extreme, then, error of prediction is smaller and test validity higher and at the other error is larger and test validity lower. Consequently the weight a test carries in prediction varies from individual to individual. Ghiselli's (1961) demonstration that two tests can be differentiated in terms of the accuracy with which they predict a criterion for a given individual is evidence of this effect. In effect Ghiselli weighted one test 1 and the other 0 for some individuals and the reverse for the remaining. Applying the optimally predicted pattern of weights for each individual accounted for a greater proportion of criterion variance. Pursuant to this position validity coefficients are average descriptions of predictive accuracy and multiple regression weights indicators of the average relative importance of the different predictors.

With respect to validity, the function of the moderator is to predict for a given individual the weight a test carries in determining criterion performance. It is not necessary that the moderator account for all criterion variance unpredicted by the tests, since some of this variance can be due to unreliability and the rest to unmeasured but important traits. The individuals' weights might be unrelated both to their criterion and test scores, or related to one or both. But nothing in this concept indicates what such correlations should be. Perhaps the correlations between the weights and the criterion and test differ from situation to situation.

Moderators are most attractive since they promise significant improvements in reliability

and especially in predictive validity. However, that other subtle variable, the suppressor, also promises much in adding to prediction but in practice seldom makes much of a contribution nor holds up well from sample to sample. Hence some counsel of caution might be in order. It is quite possible that the time and effort required to develop moderators might be more fruitfully spent in seeking improvements in reliability and validity of the sort that follow from classic psychometric theory (Ghiselli, 1960). Furthermore, since the indications are that moderators are rather specific it might be that they, like suppressors, do not hold up well from sample to sample.

REFERENCES

- ABELSON, R. P. Sex differences in predictability of college grades. *Educ. psychol. Measmt.*, 1952, **12**, 638-644.
- BERDIE, R. F. Intra-individual variability and predictability. *Educ. psychol. Measmt.*, 1961, **21**, 663-676.
- FISK, D. W. The constraints of intra-individual variability in test response. *Educ. psychol. Measmt.*, 1957, **17**, 317-337. (a)
- FISK, D. W. An intensive study of variability scores. *Educ. psychol. Measmt.*, 1957, **17**, 453-465. (b)
- FISK, D. W., & RICE, L. Intra-individual response variability. *Psychol. Bull.*, 1955, **52**, 217-250.
- FREDERICKSEN, N., & GILBERT, A. C. F. Replication of differential predictability. *Educ. psychol. Measmt.*, 1960, **10**, 759-767.
- FREDERICKSEN, N., & MELVILLE, S. D. Differential predictability in the use of test scores. *Educ. psychol. Measmt.*, 1954, **14**, 647-656.
- GHISELLI, E. E. Differentiation of individuals in terms of their predictability. *J. appl. Psychol.*, 1956, **40**, 374-377.
- GHISELLI, E. E. Differentiation of tests in terms of the accuracy with which they predict for a given individual. *Educ. psychol. Measmt.*, 1960, **20**, 615-684. (a)
- GHISELLI, E. E. The prediction of predictability. *Educ. psychol. Measmt.*, 1960, **20**, 3-8. (b)
- HOLZMAN, W. H., BROWN, W. F., & FARQUHAR, W. G. The survey of study habits and attitudes: A new instrument for the prediction of academic success. *Educ. psychol. Measmt.*, 1954, **14**, 726-732.
- HOYT, D. P., & NORMAN, W. T. Adjustment and academic predictability. *J. counsel. Psychol.*, 1954, **1**, 96-99.
- KAHNEMAN, D., & GHISELLI, E. E. Validity and nonlinear heteroscedastic models. *Personnel Psychol.*, 1962, **15**, 1-11.
- LEE, M. C. Interactions, configurations, and non-additive models. *Educ. psychol. Measmt.*, 1961, **21**, 797-805.
- SAUNDERS, D. R. Moderator variables in prediction. *Educ. psychol. Measmt.*, 1956, **16**, 209-222.
- STAGNER, R. The relation of personality to academic aptitude and achievement. *J. educ. Res.*, 1933, **26**, 648-660.

(Early publication received October 1, 1962)

DEVELOPMENT OF A FORCED-CHOICE RATING SCALE FOR ENGINEER EVALUATION¹

J. RICHARD LEPKOWSKI

International Telephone and Telegraph Federal Laboratories, Nutley, New Jersey

A technique alternative to the conventional ratings of engineers by their supervisors was studied. A 20-triad forced-choice rating scale was constructed. 33 engineers were rated by their supervisors using this device. The reliability of these ratings was .90. An item analysis showed 19 of the 20 triads to have strong discriminating power between high and low scorers. The same Ss were also rated in 8 different areas on a 4-point scale. The reliability of the 2nd rating scale was .87. The 2 scales correlated .73 with each other. These findings support previous research concerned with the more general applicability of the forced-choice technique for the determination of criterion scores.

This study investigated the value of a relatively bias-free technique in assessing the productive functioning of professional engineering personnel. Using a revision of the forced-choice rating scales developed by Stolz (1958) in a study of research productivity for physical science research workers, Cotton and Stolz (1960) investigated the value of this approach in describing the research behavior of electronics engineers. The present evaluation of a forced-choice rating scale for the appraisal of engineering personnel somewhat parallels the latter study.

PROCEDURE

Thirty-three engineers of a large electronics research laboratory were rated by their supervisors using two different evaluative devices. Seven supervisors rated from 1 to 14 different subjects. The age range of the subjects was from 22 to 41 years (mean 31.6) and total job experienced from 15 to 201 months (mean 86.6).

The Forced-Choice Rating Scale (FCS) used in this study was formed by grouping 60 selected items into 20 triads. The items were obtained from a list of 250 factor analyzed descriptive phrases concerning productive and nonproductive research workers reported in a study by Stolz (1959). The triads were constructed by selecting items highly associated with productivity and combining each with an item of moderate association and one of low or negative association. Where possible, items describing the same general characteristic (e.g., motivation, self-direction, drive, self-control, social interest) were combined.

¹ This paper reports exploratory research conducted by the Human Factors Group, Avionics Systems Laboratory, and does not reflect current personnel practices at International Telephone and Telegraph Federal Laboratories.

Supervisors were asked to check which item of a triad was most descriptive of a subject and which was least descriptive. A simple scoring system yielded triad scores of 0 to 4 and a total score (maximum 80). Reliability of the scale was determined by the split-half (odd-even) method.

The scores of the FCS were correlated with the scores of one other rating instrument using the same subjects and raters. The second measure, the Merit Appraisal Rating Scale (MAS), is a four-point rating scale providing differential weighting factors for each of four levels of engineers in eight specific areas.

The reliability of the MAS was calculated by correlating the sum of the first four ratings with the sum of the last four ratings. The content of the areas rated in each half made the analysis reasonable; each half contained two areas concerned with personal attributes and two concerned with work characteristics. The areas are: Capacity for Growth, Engineering Breadth, Engineering Judgment, and Drive *versus* Cooperation, Creativity/Ingenuity, Quality and Quantity of Work, and Leadership.

RESULTS

The FCS scores for 33 subjects ranged from 12 to 74 with a mean of 50.33, a standard deviation of 15.14, and a median of 50.

To establish the effectiveness of each triad in discriminating between high and low scorers, all triads were rescored to produce a value of either 0 or 1 and point biserial correlation coefficients were determined. Of the 20 triads, 19 produced validity indexes between .31 and .69 and 1 equaled .00.

The corrected split-half coefficients of reliability were .90 and .87 for the FCS and MAS, respectively.

The intercorrelation coefficients between the total scores of the FCS and MAS was .73 which is significant beyond the 1% level.

DISCUSSION

Since an objective criteria for evaluating the concurrent or predictive validity of each of these measures was lacking, some indication of what was being measured was gained by examining the "face" validity of each. This examination revealed that the MAS has items designed to evaluate such factors as technical knowledge, creativity, ingenuity, quality and quantity of work, resourcefulness, and ability to lead and get along with others. These qualities are generally accepted as being among the more important characteristics required of engineers working in a cooperative setting. In this sense, this scale appears to reflect good face validity.

The FCS, as noted previously, was developed from the results of earlier research. Stolz and his associates, using the Critical Incident Technique, obtained several hundred statements from supervisors of research personnel working in the physical sciences. These statements describe the most productive, least productive, and most creative men in terms of research behavior. An analysis of the collected statements resulted in a selection of

particular items to comprise the FCS. Hence, the validity of this scale is dependent upon the relation between weighted descriptions of engineering personnel and their professional activity output. Thus, it would be expected that the most productive engineers would receive the higher scores since the statements descriptive of them are more heavily weighted.

It would appear that both the FCS and the MAS have high enough reliability and intercorrelation to allow their use as checks of rater reliability. These findings also agree with the Cotton and Stolz study in that the original list of descriptive phrases concerning research productivity of physical science research workers can be used to develop a technique for evaluating productivity in a related research activity.

REFERENCES

- COTTON, J., & STOLZ, R. E. The general applicability of a scale for rating research productivity. *J. appl. Psychol.*, 1960, **44**, 276-277.
STOLZ, R. E. Development of a criterion of research productivity. *J. appl. Psychol.*, 1958, **42**, 308-310.

(Received January 2, 1962)

FACTORS AFFECTING EMPLOYEE SELECTION IN TWO CULTURES¹

HARRY C. TRIANDIS

University of Illinois

4 samples—100 Illinois and 100 Greek students, 32 Illinois and 20 Greek personnel directors (PDs)—were asked to respond to a structured questionnaire which permitted the computation of the relative weights that would be given to various characteristics (competence, age, sex, race, religion, sociability, and wealth) by these people, if they were hiring employees for various levels of jobs in the accounting and finance department of a company. The responses of the 4 samples were similar. However, the American PDs gave more weight to race and the Greek PDs more weight to age than did the other samples. The students differed significantly from the PDs; in both cultures, the students gave larger weights to competence than did the personnel directors.

It seems quite probable to this writer that personnel selection decisions are based on multiple interacting considerations. When a manager considers the hiring of a particular individual he considers several of his characteristics, such as his intelligence, education, background, previous experience, etc. Exactly how he weighs these characteristics is not clear, but it is probable that he uses some kind of implicit formula of the form: job success is a function of $X_1, X_2 \dots X_n$, with appropriate weights attached to each of the characteristics and their interactions. The present paper describes a method that permits a researcher to obtain the weights used by a manager in reaching such a decision, and illustrates this method with data from samples of personnel directors and students from two cultures.

THEORETICAL CONSIDERATIONS

It is assumed that each manager has a list of characteristics which he considers as likely to lead to job effectiveness and another list which he considers likely to lead to job failure. It is probable that the list of these characteristics differs from manager to man-

ager, and that the weights given to the negative characteristics are greater than the weights given to the positive² (Springbett, 1958). Some of the characteristics (e.g., race, sex, religion) are probably irrelevant to job success, but nevertheless may have a very substantial influence on the decision process.

The *effective* weights applied by a manager are not the *perceived* weights, but the perceived weights multiplied by the standard deviations of the distributions of the characteristics in question. For example, consider the simple situation in which a personnel director uses the following equation:

$$P = b_1X_1 + b_2X_2 + b_3X_3 + b_4X_4 + b_5$$

where P is the perceived probability that the employee will succeed on the job; X_1, X_2, X_3 , and X_4 correspond to certain characteristics of the employee and have the value +1 when the positive aspect of the characteristic is present (e.g., competent, male, white, same religion as I), and the value -1 when the negative aspect of the characteristic is present (e.g., barely competent, female, Negro, different religion than I); and the b s are weights applied to each of these characteristics. The *effective* equation, which determines the probability of exclusion of a

¹ This project was made possible through the facilities of the Institute of Labor and Industrial Relations of the University of Illinois. The trip to Greece was supported, in part, by grants from the United States Public Health Service and the Social Science Research Council for a related project (Triandis & Triandis, 1962) and for travel to Europe. Milton Derber and Lloyd G. Humphreys made most useful criticisms of an earlier version of this manuscript.

² This is understandable. The personnel director is much more likely to hear about his mistakes than about his successes. If he hires an alcoholic, for instance, he is sure to be asked, "How did you miss this?" while if his choice is good, people will simply assume that he is doing what he is supposed to be doing.

person from the job, is not the one above, but the equation which is obtained by multiplying the b weights by the corresponding standard deviations of the distributions of the characteristics. In the case of dichotomous characteristics, these standard deviations can be estimated from the usual $\sqrt{p \cdot q}$ formula. For instance, for woman-man the σ is $\sqrt{.50 \times .50}$, or .50, while for Negro-white it is $\sqrt{.10 \times .90} = .30$, for Jewish-Christian $\sqrt{.05 \times .95} = .22$, and for competent-incompetent it would depend on the difficulty of the job. For very difficult jobs, such that only one person in 10,000 can perform them adequately, it would be very small ($\sqrt{.0001 \times .9999} = .01$). Thus, for very difficult jobs the *effective* weight for sex may be equal to 50 times the weight for competence, *even when the manager perceives these characteristics to be of equal weight!*

The method used in the present study involves complex stimuli arranged in a factorial design. Each stimulus includes four characteristics, such as the sex, age, competence, race, etc. of a hypothetical job applicant, and the personnel director (PD) is asked to make a judgment on a seven-point scale ranging from "I would strongly recommend" to "I would strongly oppose the hiring" of this applicant. Since the stimulus characteristics are arranged in a factorial design, it is possible to compute analyses of variance indicating: (a) how much variance is controlled by each characteristic, (b) whether this variance is reliably greater than the error variance (i.e., whether the subject pays attention to a particular characteristic in making his judgment), and (c) whether the subject perceives interactions between the characteristics (e.g., Negro women may be perceived as being particularly unsuitable for the job while white women are particularly suitable). The method was first used by Triandis and Triandis (1960) and has been used again in a number of other studies (Rickard, 1962; Triandis & Triandis, 1962).

One of the advantages of the method is that the estimates of the percentage of variance controlled by each of the characteristics are proportional to the beta weights of a regression equation, such as the one shown above. This is so because the design is com-

pletely balanced. The X s in the above-mentioned equation are in standard score form. The beta weights are obtained from the cross products of the X s and the P s of the equation. These weights are proportional to the difference between the sum of the P s obtained when the X s are equal to +1 and the sum of the P s obtained when the X s are equal to -1. Since in the analysis of variance this same difference determines the magnitude of the sum of squares, the weights obtained from the analysis of variance are proportional to the regression weights. (Note: this is only true when a given factor has only one degree of freedom.)

METHOD

Samples

Four samples were used: (a) 100 University of Athens students,³ (b) 100 University of Illinois students,³ (c) 20 Greek personnel directors, (d) 32 Illinois personnel directors.

Characteristics of the personnel director samples. The Greek personnel directors were picked from the largest industry of each kind that was available in the Athens metropolitan area. If there were examples of both publicly and privately controlled companies in a given kind of industry, personnel managers from two industries of that kind (one public, the other private) were interviewed. Thus, the sample included the following enterprises: the leading chemical processing company, the leading ceramics manufacturing company, the leading textile company, the leading armaments company, the leading tobacco manufacturing company, the most important railway, the electrification company of Greece, the leading airline, the leading distillery, the leading utilities in the Athens area, the leading shipyard, the leading chocolate maker, one of the major makers of appliances, one of the major soap makers, one of the major cement makers, and the leading department store. The sample was worked out with the help of the Greek Productivity Center, the Ministry of Labor, and the Greek Association of Manufacturers. The industries had from 300 to 8,000 employees (the median was 700). Forty-two percent of the sample was in top management (the managing director makes the personnel decisions), 53% was in middle management, and 5% in lower management. Forty-two percent was in the 50-60 age group, 42% in the 40-50, and 16% in the 30-40 age group. Thirty-

³ The University of Athens students were asked to fill in the questionnaire while they were studying in the library of the student union during the summer of 1960. The University of Illinois students were in their first month of the course in introductory psychology. The samples were rather similar in most socioeconomic characteristics (Triandis & Triandis, 1962).

two percent of the respondents was classified by the writer as belonging to the Greek upper class, 10% to the upper middle, and 58% to the lower middle class. Fifty-three percent had completed their university education.

The Illinois personnel directors were approached by means of a mail questionnaire. The names of 106 personnel directors who had attended the 1957 personnel management conference held at the University of Illinois were obtained from the registration list. Questionnaires were mailed to all these persons but only 32 returned them. They stated that their companies had a range from 3 to 35,000 employees with a median of 600 (more dispersed than the Greek sample but otherwise similar). Forty-four percent said they were in top management, 50% was in middle management, and 6% in lower management (surprisingly similar to the Greek sample). Thirty-one percent was in the 50-60 age group, 22% in the 40-50, and 47% in the 30-40 group. (They were much younger than the Greeks.) Their background was such that none was considered to belong to the American upper class, 6% belonged to the upper middle, 53% to the lower middle, and 41% to the working class (a striking difference from the Greek sample which suggests the difficulty of upward mobility in Greece). Sixty-three percent had completed their university education.

The Greek sample included older persons, more who were upper class, and individuals with less formal education than the Illinois sample.

Research Instrument

The basic research instrument was exactly the same for all four samples. It consisted of a questionnaire which included 32 stimuli in the form of persons with specified characteristics, under each of which was a seven-point scale ranging from "I strongly oppose" to "I strongly recommend" the hiring of this individual. The stimuli may be seen in Table 1.

All samples answered the questionnaire by themselves. The Greek personnel directors, however, answered it in the course of a lengthy interview which explored, among other things, the additional factors which influence personnel decisions in Greece. Such data are not available from the Illinois sample. On the other hand, the Illinois sample answered the questionnaire three times: once while thinking of the job of section manager of the accounting department, the next time while thinking of the manager of the district finance office, and the last time while thinking of the company comptroller's job. The Greeks answered the questionnaire only once and were told to think of the job of the highest official in the accounts and finances department—a job that should correspond to the company comptroller's job. Some of the Greeks stubbornly precluded any possibility of hiring the comptroller from the outside. They insisted that all promotions were made from within the company. This made it necessary to let them answer the questionnaire with the lowest level of

management in mind. Both of the student samples were asked to think of the job of the manager of the finance and accounts department.

The Illinois sample was also asked a large number of personality and attitude scale questions. A more thorough discussion of the results of an analysis of these items can be found in Triandis and Triandis (1962).

Definitions of terms. Certain terms used in the questionnaire were defined for the respondents. The "highly competent" applicant was defined as one whose education, recommendations, and objectively obtained aptitude scores were highly favorable as far as the particular job was concerned. The words "barely competent" indicated that the objective evidence was consistent, but barely favorable. The word "sociable" was defined as indicating a warm, friendly, and outgoing personality. "Unsociable" was defined as indicating a cold, reserved, and somewhat unfriendly approach to human relationships.

RESULTS

Table 1 shows the means of the scores obtained from the four samples of subjects. A visual inspection suggests that the data are highly correlated. Statistical treatment yields the following Pearson correlations:

$r_{\text{American students-American personnel directors}} = .95$
(based on $N = 32$), $r_{\text{Greek students-Greek personnel directors}} = .77$, $r_{\text{American and Greek students}} = .94$,
and $r_{\text{American and Greek personnel directors}} = .83$.

Thus, the only sample that gives results that are slightly different from the other three samples is the sample of Greek personnel directors. However, if it is recalled that there was a difference in method in the case of this sample (the questionnaire was answered in the course of an interview and was *not* anonymous while in the case of the other three samples *it was* anonymous), then the correlations obtained between the results of this sample and the others are not surprising. These correlations suggest a high reliability for the method used. They also suggest that judgments of employability are highly consistent across cultures and experiences with employee selection (student versus PD).

We also considered whether the *differences* between the Greek and American personnel director scores were correlated with the differences between the Greek and American students' scores. The signs of those differences are rather consistent ($p < .001$ by chi square) which means that if the Greek PDs are willing to employ a particular applicant

TABLE 1
MEAN SCORES OBTAINED FROM THE FOUR SAMPLES

Source (characteristics)	American PDs Judging :			American students judging :	Greek PDs judging :	Greek students judging :
	Department Manager	District Manager	Comptroller	Comptroller	Manager—Accounting and Finance Division	
1. Highly competent 55-year-old male Negro	4.90	4.80	4.90	2.99	5.75	3.34
2. High competent male Jew, unsociable, from rich home	5.45	5.50	4.65	4.08	3.90	4.40
3. Barely competent 30-year-old Negro male	6.72	6.50	6.66	6.03	5.95	5.11
4. Sociable male Jew, poor home, barely competent	6.04	6.42	6.29	5.46	5.35	4.53
5. Highly competent white male, 30-year-old	1.56	1.97	2.71	1.18	1.25	1.82
6. Highly competent unsociable Christian, rich home	4.70	4.22	3.90	3.06	3.25	4.38
7. White woman, 30-year-old, barely competent	4.35	6.50	6.58	5.76	5.25	5.54
8. Barely competent Christian, sociable, poor	5.82	6.10	6.04	4.82	6.63	3.65
9. White woman, 55-year-old, highly competent	4.20	4.04	4.71	2.43	6.20	3.62
10. Highly competent male Jew, sociable, rich	2.96	2.38	2.29	1.78	2.70	2.80
11. Barely competent 55-year-old male Negro	6.72	6.78	6.60	6.22	6.85	5.43
12. Highly competent, very sociable Christian, poor	1.75	1.42	1.38	1.20	1.40	1.35
13. Barely competent 30-year-old Negro woman	6.78	6.76	6.85	6.19	5.40	5.12
14. Male Jew, sociable, competent, rich	6.05	6.20	6.24	5.69	5.85	4.81
15. Highly competent Christian, sociable, rich	2.87	3.48	4.68	1.49	2.90	3.10
16. Barely competent Christian, sociable, rich	5.85	5.74	6.10	5.20	5.35	4.12
17. Negro, barely competent 55-year-old woman	6.80	6.60	6.80	6.36	7.00	5.83
18. Highly competent Christian, sociable, rich	1.87	1.69	1.55	1.29	1.95	2.03
19. Highly competent 55-year-old white male	2.90	2.77	2.23	2.02	5.30	2.12
20. Poor, unsociable, highly competent Christian	4.60	4.32	4.06	2.97	2.25	3.08
21. 30-year-old highly competent Negro woman	4.78	5.20	5.55	2.70	4.55	3.46
22. Rich, barely competent unsociable, Jew	6.72	6.72	6.55	6.31	5.70	5.46
23. Highly competent 55-year-old Negro woman	5.56	5.34	5.80	3.51	6.60	3.94
24. Highly competent sociable Jew, poor	2.28	2.16	2.16	1.75	2.55	2.42
25. Negro, 30-year-old, highly competent male	3.75	4.10	4.70	2.17	3.90	2.56
26. Barely competent, unsociable, Christian, rich	6.60	6.48	6.58	5.96	5.80	5.30

Note.—1 = strongly recommend ; 7 = strongly oppose.

Table 1—Continued

Source (characteristics)	American PDs Judging:			American students judging:	Greek PDs judging:	Greek students judging:
	Department Manager	District Manager	Comptroller	Comptroller	Manager—Accounting and Finance Division	
27. Barely competent, white, 30-year-old male	6.18	6.30	6.35	5.50	4.90	4.67
28. Unsociable, male, Jew, highly competent, poor	5.00	4.86	4.45	3.03	3.30	3.90
29. White woman, 55-year-old, barely competent	6.63	6.58	6.78	5.95	6.85	5.59
30. Poor, unsociable, Jew, barely competent	6.68	6.74	6.78	6.35	6.30	5.59
31. White 55-year-old man, barely competent	6.72	6.68	6.68	6.06	7.00	5.44
32. Christian, barely competent, unsociable, poor	6.81	6.61	6.78	6.10	5.96	5.27

while the Illinois PDs are not quite as keen, then the Greek students are also more willing to employ the applicant than are their American counterparts, and vice versa. In other words, there is a culturally determined tendency to respond to the questionnaire. However, when the *actual* differences (rather than their signs alone) are considered we find that they are not significantly correlated ($r = .11$, not significant).

The correlations mentioned above also suggest that the variance due to anonymity is greater than the variance due to culture or occupation of respondents.

To establish the statistical significance of these correlations David's (1938) tables, which are reproduced in Pearson and Hartley (1954), were used, since the correlations are based on means. For the number of observations in question (32), the range of .57–.93 is significant at the 1% level. The obtained correlations are well within this range.

Against the background of similarities suggested by these correlations it is possible to evaluate more clearly the differences in the weights placed on the various characteristics used. Table 2 presents the sums of squares and percentage of variances obtained when the scores of Table 1 were subjected to analyses of variance. This statistical procedure permits the computation of the weights (percentage of variances) given to the characteristics. It is also possible to learn whether a characteristic con-

trolled a statistically significant amount of variance. If it did not do so, the subjects paid no attention to that characteristic.

A summary of the percentage of variances (weights) given to the various characteristics is shown in Table 3.

A final attempt to assess the relative importance of culture was an analysis of variance in which the sources of variance were the following: competence, age, race, sex (of stimulus person judged), occupation (student or PD), and culture (American or Greek) of respondent. This analysis is summarized in Table 4.

Table 1 shows that a highly competent, white, 30-year-old male (Stimulus Number 5) is enthusiastically recommended for any of the three jobs (though the Illinois PDs are not quite as enthusiastic for the top job—presumably because they consider him a bit too young). Similarly, Stimuli Numbers 12 and 13 are equally acceptable. A single change in the characteristics, a change in race from white to Negro (Stimulus Number 25), causes a most considerable tempering of the enthusiasm. In fact, since 4 is the midpoint of the scale, it is clear that the Illinois PDs reject this stimulus for the jobs that are not at the very bottom of the management hierarchy. Even the Greek PDs are on the verge of rejecting him. On the other hand, the students in both samples are still willing to hire him. The addition of just one more negative characteristic (woman—see

TABLE 2
ANALYSES OF VARIANCE RESULTS BASED ON DATA OF TABLE 1

Source (char- acteristic)	American PDs judging:						American students judging:			Greek PDs judging:			Greek students judging:		
	Department Manager			District Manager			Comptroller			Manager—Accounting and Finance Division					
	SS	%	F	SS	%	F	SS	%	F	SS	%	F	SS	%	F
Competence	25.96	60.5	52.0***	27.54	73.8	69.4***	20.3	63.8	41.4***	10.16	26.2	18.8**	21.80	82.0	170.0***
Race	7.02	16.0	12.3**	3.75	10.0	9.4*	3.1	9.7	6.3*	2.52	6.5	4.7	0.56	2.1	4.6
Age	3.46	8.1	6.1*	0.49	1.3	—	0.0	0.0	0.0	19.03	49.5	35.6***	1.01	3.8	8.0*
Sex	0.40	0.9	—	1.33	3.5	3.34	3.0	9.4	6.1*	0.92	2.4	1.7	1.97	7.4	16.2**
Interaction	6.14	14.0	—	4.38	11.7	—	5.4	17.5	—	5.9	15.3	—	1.32	5.0	—
Total	42.98			37.49			31.8			38.5			26.66		
Competence	30.14	63.1	35.2***	37.42	67.0	38.6***	45.29	77.0	119.0***	38.55	82.0	71.4***	12.90	51.5	100.0***
Religion	0.63	1.3	—	1.22	2.2	—	0.57	1.0	—	0.59	1.3	—	1.40	5.6	10.8**
Sociability	12.14	25.5	14.1**	11.14	20.0	11.5**	8.56	14.6	22.5***	1.36	2.9	2.5	8.51	34.0	65.0***
Wealth	0.10	0.2	—	0.00	0.0	—	0.00	0.0	—	0.36	0.8	—	0.78	3.1	6.0*
Interactions	4.52	9.5	—	5.93	10.7	—	4.14	7.1	—	5.94	12.6	—	1.40	—	—
Total	47.53			55.72			58.56			46.90			24.99		

* $p < .05$.
** $p < .01$.
*** $p < .001$.

Stimulus Number 21) is enough to swing all PDs into a firm opposition, for no matter what managerial job (though the opposition increases with job level), and substantially tempers the enthusiasm of the students (the Greeks are on the verge of rejection now). One more negative characteristic—55 years old (see Number 23)—and even the Illinois students are on the verge of rejection. Competence is important for all groups. A barely competent white 30-year-old (Number 27) is rejected by all samples, particularly the Americans.

Since the stimulus characteristics were arranged so that an analysis of variance could be computed, it is possible to examine the statistical significance of the various characteristics. Table 2 shows that competence, race, and sociability are important characteristics for all the jobs judged by the Illinois PDs. The Illinois students consider competence, religion, and sociability, and do not stress race as much as the Illinois PDs. The PDs pay attention to age at the lowest job level (they do not want to hire a 55-year-old section manager) but ignore this characteristic at the higher levels. The Greek PDs also pay attention to competence and race, but pay little attention to sociability. On the other hand, they are greatly concerned with the age of the applicant. The Greek students, finally, pay attention to all the characteristics. It is also worthy of notice that the Illinois PDs pay attention to sex only for the highest job—sex gains in importance as the job level increases.

Table 3 was obtained by assuming that the weights obtained from a given sample of respondents for competence, in the two analyses of variance involving those respondents (Table 2), can be used as standards for the scaling of the other weights in one of these analyses, so that a summary statement of weights for all seven factors can be obtained. It shows that competence and sex increased in importance as the PDs considered higher jobs; and this was done at the expense of race, sociability, and age. The Illinois respondents were much more concerned with competence and race than the Greeks. The American greater stress on competence, as compared to the Greek, sug-

TABLE 3
SUMMARY OF PERCENTAGE OF VARIANCES FROM TABLE 2

Source (characteristic)	American PDs judging:			American students judging:	Greek PDs judging:	Greek students judging:
	Department Manager	District Manager	Comptroller	Comptroller	Manager—Accounting and Finance Division	
Competence	45.0	54.0	54.0	70.0	24.8	46.3
Race	12.1	7.3	8.2	3.0	6.1	1.2
Age	6.0	1.0	0.0	1.8	46.0	2.1
Sex	0.7	2.6	8.0	0.4	2.1	4.2
Religion	0.9	1.7	0.7	1.5	0.0	5.0
Sociability	18.0	16.2	10.2	11.1	1.0	30.6
Wealth	0.1	0.0	0.0	0.2	0.2	2.8
Interactions	17.2	17.3	18.9	12.1	18.7	7.8

gests that the Protestant ethic (Weber, 1930) is still fairly dominant in the American scene in spite of suggestions to the contrary (Riesman, 1953).

Although race was a more important factor in Illinois than in Greece, even in Greece there was opposition to the hiring of Negroes. Various "justifications" for this bias were presented by the personnel directors. Difficulties in social adjustment in the work environment; high unemployment rates among white workers; and in some cases, biological inferiority of Negroes were given as reasons for the bias.

Greek PDs were also less concerned with the sex of the applicant, for the jobs that were below the top of the managerial hierarchy, than were American PDs. The greater

acceptance of women in the lower and middle ranks was typical of most respondents, though a significant minority had a strong bias against hiring women for positions of responsibility.

The Greek PDs were much more concerned than American PDs with the age of the applicant. They were unanimous about not hiring the 55-year-old applicants. This bias was also present in the American sample, but in Greece it was emphasized by the particular provisions of the laws governing pensions and severance pay, which require relatively large compensation at the time of termination of employment (see Triandis, 1961, for details).

One factor which produces significant differences between the American and Greek hiring patterns is the extent of government specification of the characteristics of applicants in Greece. Many firms, particularly the semipublic firms such as the utilities and the railways, have government-approved "constitutions" which specify that those hired must be Greek citizens, Greek Orthodox, and under 35 years old. This practice essentially institutionalizes discrimination because of race, religion, nationality, and age in a manner that is unknown in the United States.

Finally, Table 4 shows that, considering all four samples together, there was no difference in the way the subjects responded that can be attributed to culture (i.e., in a general sense both cultures answered the questionnaire similarly), yet there were sig-

TABLE 4

ANALYSIS OF VARIANCE WITH CHARACTERISTICS OF RESPONDENTS AS A SOURCE OF VARIANCE

Source	df	Variance	F
Competence	1	974,400	159.0***
Age	1	109,470	17.9***
Race	1	76,900	12.5***
Sex	1	54,460	8.9**
Occupation of respondents	1	263,000	43.0***
Culture of respondents	1	107	
Interactions	57	6,140	

** $p < .01$.

*** $p < .001$.

nificant differences between personnel directors and students. Furthermore, the four samples, as a whole, considered all four characteristics (competence, age, race, and sex) important in the selection process.

The Greek hiring pattern is described in more detail in Triandis (1961). This report also discusses individual differences between PDs in the two cultures.

DISCUSSION

It is clear, from the data presented above, that personnel directors in both cultures pay attention to characteristics such as older, Negro, woman, etc., which are probably not negatively correlated with job success. The weights used for these allegedly negative characteristics are relatively small, but when these weights are multiplied by the standard deviations they become quite important. Under certain conditions, these weights may bar most of the qualified candidates with these characteristics from the job. The problem is not limited to the characteristics studied in the present project; for instance, many personnel directors and businessmen require that an applicant be a college graduate. However, a college education is *not* always necessary for success on the job; it would be much more reasonable to determine this matter empirically. A research program is needed that would examine whether or not the characteristics which personnel directors consider important are, in fact, relevant to job success.

The use of irrelevant characteristics in the hiring process can be particularly unfortunate because it increases the selection ratio (number hired over number considered for the job). This is especially crucial in regard to the top jobs in business and industry, where the supply of talent is limited.

The method presented in the present paper and illustrated with data from two cultures is especially promising since it permits a description of what personnel directors consider important in employment. It could be adapted in many ways. For instance, it could be used with specific test, interest, and personality inventory scores for various kinds of jobs with various personnel directors. If differences were obtained between the weights given by the personnel directors to each of the characteristics, it would be possible to

ask if these differences could be accounted for by the training, background, personality, etc., of the personnel directors. Some personnel directors, with their particular sets of weights, may be making particularly effective personnel decisions. It would be possible to ascertain under what conditions (jobs, industries) certain sets of weights lead to the most effective decisions.

An illustration of the usefulness of this procedure with personality characteristics of personnel directors is provided by Rickard (1962). Rickard used the method with personnel directors and school administrators, with reference to positions as accountants and teachers. The stimuli varied in disability (epileptic, person discharged from prison, ex-mental patient, deaf person, person confined to a wheelchair, and person discharged from a tuberculosis sanitarium). The rankings of the importance of these disabilities were extremely reliable. Some personnel directors gave larger weights to the disability than did others, and those giving these larger weights were found to be higher on the California F Scale and the Rokeach Dogmatism Scale. Some other aspects of that study are reported in Rickard, Triandis, and Patterson (1963).

REFERENCES

- DAVID, FLORENCE N. *Tables of the ordinates and probability integral of the distribution of correlation coefficients in small samples*. Cambridge: Cambridge Press for Biometrika Trustees, 1938.
- PEARSON, E. S., & HARTLEY, H. O. *Biometrika tables for statisticians*. Cambridge: Cambridge Press, 1954.
- RICKARD, T. E., TRIANDIS, H. C., & PATTERSON, C. H. Indices of employer prejudice toward disabled applicants. *J. appl. Psychol.*, 1963, **47**, 52-55.
- RIESMAN, D., GLAZER, N., & DENNEY, R. *The lonely crowd*. New York: Anchor, 1953.
- SPRINGBETT, B. M. Factors affecting the final decision in the employment interview. *Canad. J. Psychol.*, 1958, **12**, 13-22.
- TRIANDIS, H. C. Factors affecting employee selection in two cultures. Author, 1961. (Mimeo)
- TRIANDIS, H. C., & TRIANDIS, L. M. Race, social class, religion, and nationality as determinants of social distance. *J. abnorm. soc. Psychol.*, 1960, **61**, 110-118.
- TRIANDIS, H. C., & TRIANDIS, L. M. Some determinants of social distance in two cultures. *Psychol. Monogr.*, 1962, **76**(21, Whole No. 540).
- WEBER, M. *The Protestant ethic and the spirit of capitalism*. London: Allen & Unwin, 1930.

(Received March 12, 1962)

PERSONAL ADJUSTMENT AND ACADEMIC PREDICTABILITY AMONG COLLEGE FRESHMEN

L. BRYCE ANDERSEN AND PATRICIA A. SPENCER

University of Minnesota

The objective of this study was to investigate whether the prediction of academic achievement is influenced by personal emotional adjustment as found by Hoyt and Norman in 1954. Samples consisted of 1,465 arts college freshmen and 620 engineering college freshmen from which were selected three "adjustment" groups (normal, one-peak, and maladjusted) according to arbitrary cut-off points on the clinical scale of the MMPI. Correlations between grade point average (GPA) and predictor variables were determined. For the arts college, no significant differences were found between the adjustment groups for the correlation of GPA with either of the predictors high school rank or the Minnesota Scholastic Aptitude Test. 11 predictor variables were used for the engineering freshmen, only one of which yielded statistically significant differences between the adjustment groups. In contrast to the findings of Hoyt and Norman, it was concluded that the prediction of academic achievement is not influenced by personal adjustment.

An important problem in predicting academic success concerns the effect of emotional and social maladjustment on predictability. Do students with personal emotional problems "fool" predictions made by academic ability tests?

This problem was studied by Hoyt and Norman (1954) who investigated the hypothesis that academic ability tests will predict grades less well for a maladjusted group than for a well-adjusted group. They studied 328 University of Minnesota freshman men enrolled in the College of Science, Literature, and the Arts in 1951-52 and 1952-53. Three "adjustment" groups were selected based on arbitrary cutoff points on the clinical scales of the Minnesota Multiphasic Personality Inventory (MMPI). A "normal" group consisted of students with no *T* scores above 60 on the clinical scales. A "one-peak" group included those students who had one *T* score of 70 or above, excluding the *Mf* scale. A "maladjusted" group showed two or more *T* scores of 70 or above, excluding *Mf*. The authors assumed that the degree of disturbance is reflected by the height of the profile.

Correlations between first quarter grade point average (GPA) and two aptitude tests, the American Council on Education (ACE) Psychological Examination (1947 edition) and the Ohio Psychological Examination (Form 22), were determined. Hoyt and Nor-

man found statistically significant differences among the three groups in the correlations between the Ohio and GPA. The normal group was significantly more predictable than the one-peak and the maladjusted groups. However, the results using the ACE as a predictor were statistically not significant and were ambiguous. The ACE appeared to have approximately equal effectiveness as a predictor for all three groups. In one case, the maladjusted group appeared to be more predictable than the normal group.

The Hoyt and Norman results suggested the present study. The primary objective of this study was to investigate whether maladjustment would result in a lesser degree of academic predictability among two large samples of freshmen, as it had for the smaller sample of freshmen in the Hoyt and Norman study. The study is divided into two parts and, because the sampling and procedure vary somewhat, they will be discussed separately in the appropriate sections.

SAMPLES AND PROCEDURE

University of Minnesota freshmen from the College of Science, Literature, and the Arts (SLA) and the Institute of Technology (IT) were divided into adjustment groups on the basis of the MMPI according to the method of Hoyt and Norman. Although this numerical criterion is simple and unambiguous, it does not given any consideration to the subtleties of the MMPI as expressed by the relationship of

one score to another. The GPA was the criterion variable. At the University of Minnesota, this is based on the following scale: $A = 4.0$, $B = 3.0$, $C = 2.0$, $D = 1.0$, and $F = 0.0$.

SLA Freshmen. This part of the study was a repetition of the Hoyt and Norman study on a large sample of 1959-60 freshmen, including both men and women. Data for 1959-60 freshmen who graduated from Minnesota high schools in 1959 were available from an earlier study (Swanson & Berdie, 1961). Of these freshmen, only those with MMPI profiles were included in the present study. The sample totaled 1,465 students (691 women and 774 men). These 1,465 students were divided into three adjustment groups within each sex, according to the Hoyt and Norman method, with one exception. In this study, the *Mf* scale on the MMPI was excluded as a consideration in selecting the normal group, whereas Hoyt and Norman included it in their selection of the normal group.

The number in each adjustment group and the mean MMPI profile for each group are given in Table A.¹ In addition to the normal, one-peak, and maladjusted groups, there was a residual group consisting of 514 students (265 women and 249 men) who had no *T* score above 70, but at least one *T* score above 60 on the clinical scales. This average group was not defined or considered by Hoyt and Norman; nor is it here, since it would not have been relevant to the objective of this study. The MMPI data on the total SLA college freshman men and women were not available at the time of this study. As one would anticipate, Table A shows that the mean MMPI *T* scores for both men and women were the lowest for the normal group and the highest for the maladjusted groups due to the selection criterion.

Two predictors were considered: high school rank (HSR) obtained at the end of the junior year in high school, and the Minnesota Scholastic Aptitude Test (MSAT) administered in January and February of the junior year in high school. The GPA for the entire year is the criterion variable, whereas Hoyt and Norman used first quarter GPA. Bivariate correlation coefficients were determined among GPA and the two predictor variables. The resultant correlation coefficients were then tested, using Fisher's *z*-transformation test, for significant differences among the adjustment groups and between the respective male and female adjustment groups.

IT Freshmen. This aspect of the study involved 620 1960-61 engineering college freshmen. Data for the total engineering freshman class were available from an earlier study (Swanson, 1961). These students were divided into adjustment groups follow-

ing the procedure of Hoyt and Norman. Table B¹ reports the mean MMPI profile for each adjustment group, and the number in each group, plus similar data for the total freshman class. The residual or "average" group ($N = 333$) was not considered here. The adjustment groups and the average group include 575 students, whereas the total group includes 620 students. The difference is due to students with incomplete records and to a small group of women engineering freshmen ($N < 5$) who were excluded because of the different MMPI scoring. Hence, this part of the study is based on an all-male group. Again, due to the selection criterion, the mean MMPI *T* scores range from the lowest for the normal group to the highest for the maladjusted group, as reported in Table B.

Eleven predictor variables and GPA for the first quarter were considered. The predictors were: HSR obtained at the end of the junior year in high school; MSAT administered in January and February of the junior year in high school; Co-operative English Test (CET), Form Z, lower level, administered at the same time as the MSAT; Institute of Technology Mathematics Test (IT Math.) administered in March and April of the senior year in high school; American College Testing (ACT) Program administered in November 1959 and February and April 1960, including ACT English, ACT Mathematics, ACT Social Studies, ACT Natural Sciences, and ACT Composite; College Entrance Examination Board (CEEB) Scholastic Aptitude Tests administered in August and September 1960, including the CEEB Verbal and the CEEB Mathematics. Bivariate correlation coefficients were determined among GPA and the 11 predictor variables, and then tested (Fisher's *z*-transformation test) for significant differences among the adjustment groups.

RESULTS AND DISCUSSION

SLA Freshmen. The bivariate correlation coefficients for the two predictor variables and GPA are reported in Table 1 and the means and standard deviations in Table 2. The "total" group reported is that of the earlier study by Swanson and Berdie (1961). As indicated, all the correlation coefficients for the adjustment groups are significantly different from zero at the .01 level.

Of greatest interest in this study is the comparison of the correlation coefficients for different adjustment groups. There are certain statistical procedures for comparing the groups with each other since they are independent. However, the total group contains the adjustment groups and no method is available for making comparisons here. None of the differences in correlations, for either the male or the female adjustment groups, is

¹Table A, Table B, and Table C have been deposited with the American Documentation Institute. Order Document No. 7407 from the ADI Auxiliary Publications Project, Photoduplication Service, Library of Congress; Washington 25, D. C., remitting in advance \$1.25 for each microfilm or \$1.25 for photocopies. Make checks payable to: Chief, Photoduplication Service, Library of Congress.

TABLE 1

CORRELATION COEFFICIENTS BETWEEN PREDICTORS AND GRADE POINT AVERAGE FOR SLA FRESHMEN

Group	N	HSR-GPA	MSAT-GPA
SLA men			
Total	821	.433	.371
Normal	238	.457*	.380*
One-peak	162	.489*	.384*
Maladjusted	125	.535*	.388*
SLA women			
Total	706	.528	.467
Normal	258	.579*	.454*
One-peak	103	.475*	.508*
Maladjusted	65	.450*	.586*

* Significantly different from zero at the .01 level.

statistically significant for either of the predictor variables. That is, the one-peak and the maladjusted groups are no less predictable than the normal group. Furthermore, tests between respective male and female adjustment groups show no statistically significant differences. For example, normal men and normal women are equally predictable.

Table 2 shows some statistically significant differences in means of predictors. For example, one-peak and maladjusted men score higher on the MSAT than do normal men; but for women, only the one-peak group

scores higher than the normal group. Normal women and one-peak women average higher on high school rank than do maladjusted women, whereas there are no differences for men. None of the GPA means is significantly different among either the male or the female adjustment groups, all tending to be similar and near a "C" average. This is consistent with the results thus far reported and indicates that academic achievement as measured by first year GPA is not influenced by personal emotional adjustment as based on the MMPI. None of the male or the female adjustment groups is significantly more variable than the others.

IT Freshmen. Bivariate correlation coefficients for the 11 predictor variables and first quarter grade point average are reported in Table 3 for the three adjustment groups and the total engineering freshman class. The majority of the correlation coefficients are significantly different from zero as indicated.

All the correlation coefficients have been tested to determine significant differences among the three adjustment groups. An examination of the correlations between grade point average and the various predictor variables shows little variation among the three adjustment groups. In only one case of the 11 predictor variables is the difference in GPA correlation of the normal and the maladjusted groups significant. The sole signifi-

TABLE 2

MEANS AND STANDARD DEVIATIONS OF THE PREDICTOR VARIABLES AND GRADE POINT AVERAGE FOR SLA FRESHMEN

Group	HSR		MSAT		GPA	
	M	SD	M	SD	M	SD
SLA men						
Total	67.51	21.15	41.81	11.66	— ^a	— ^a
Normal	68.19	20.61	40.63	11.73	1.94	.79
One-peak	67.41	21.70	42.89 ^b	11.43	1.95	.75
Maladjusted	67.92	20.86	44.17 ^b	11.81	1.87	.74
SLA women						
Total	78.78	18.50	44.95	12.87	— ^a	— ^a
Normal	80.13 ^c	16.76	43.98 ^d	11.81	2.23	.77
One-peak	80.57 ^c	16.89	47.73	12.97	2.25	.74
Maladjusted	71.67	20.89	43.83	12.93	2.00	.77

^a Data not available.
^b Significantly different from the equivalent value for the normal group at the .01 level.
^c Significantly different from the equivalent value for the maladjusted group at the .01 level.
^d Significantly different from the equivalent value for the one-peak group at the .05 level.

TABLE 3

CORRELATION COEFFICIENTS BETWEEN PREDICTORS AND GRADE POINT AVERAGE FOR IT FRESHMEN

Correlations	Total	Normal	One-peaked	Maladjusted
GPA and:				
HSR	.390 ^a	.373 ^a	.453 ^a	.357 ^a
MSAT	.344 ^a	.410 ^a	.177 ^a	.501 ^a
CET	.370 ^a	.377 ^a	.225	.399 ^a
IT Math.	.629 ^a	.675 ^a	.622 ^a	.693 ^a
ACT English	.361 ^a	.333 ^a	.226	.436 ^a
ACT Mathematics	.408 ^a	.497 ^a	.382 ^a	.154
ACT Social Studies	.322 ^a	.343 ^a	.143	.365 ^a
ACT Natural Sciences	.340 ^a	.389 ^a	.317 ^b	.199 ^a
ACT Composite	.455 ^a	.491 ^a	.340 ^a	.290 ^b
CEEB Verbal	.411 ^a	.436 ^a	.301 ^b	.624 ^a
CEEB Mathematics	.420 ^a	.435 ^a	.376 ^b	.424 ^a
<i>N</i>	620	112	70	60

^a Significantly different from zero at the .01 level.^b Significantly different from zero at the .05 level.

cant difference is for ACT Mathematics-GPA (.05 level). None of the GPA correlations for the one-peak group is significantly different from the corresponding values for the normal group. Curiously, two of the one-peak correlations are significantly different from those of the maladjusted group. Academic predictability among engineering freshmen is not influenced by personal adjustment.

As reported in Table C,¹ the mean values of the predictor variables among the adjustment groups are not significantly different, with one exception. This sole exception indicates that the maladjusted group scores higher on the CEEB Verbal than do the other two groups. The mean values of the criterion variable, grade point average, are not significantly different among the three adjustment groups. All have averages near a C. The variability of scores, as measured by the standard deviations, is similar for the three adjustment groups, with the few exceptions as indicated. Hence, as measured by group means and standard deviations, the three adjustment groups are remarkably similar to each other in aptitude and achievement.

CONCLUSIONS

Personal emotional adjustment, as defined by Hoyt and Norman using MMPI clinical

scale profiles, does not influence academic achievement of either arts college or engineering freshmen, as determined by their GPA. Prediction of achievement using either of the predictors for arts college freshmen, or any of the 11 predictors for engineering freshmen, is not significantly influenced by the personal adjustment of the student. A normal group is not significantly more predictable than a maladjusted group. Indeed, the three adjustment groups are remarkably similar in aptitude, achievement, and predictability.

REFERENCES

- HOYT, D. P., & NORMAN, W. T. Adjustment and academic predictability. *J. counsel. Psychol.*, 1954, 1, 96-99.
- SWANSON, E. O. Correlations of first quarter grade point averages (GPA) for fall 1960 entering freshmen in the Institute of Technology, the University of Minnesota, versus scores from the Minnesota State-Wide Testing Program, American College Testing Program (ACT), and College Entrance Examination Board (CEEB) Scholastic Aptitude Tests (*N* = 620). Student Counseling Bureau memorandum, 1961, University of Minnesota.
- SWANSON, E. O., & BERTIE, R. F. The relation of Minnesota college state-wide program test scores to first year grade point averages in Minnesota colleges and a survey of scholastic aptitude in Minnesota colleges. *U. Minn. res. Bull., Off. Dean Students*, 1961, 3(1).

(Received March 29, 1962)

PREDICTION OF ATTENDANT TENURE

EARL C. BUTTERFIELD

AND

SUE A. WARREN

Southbury Training School, Connecticut

Oregon Fairview Home, Salem

MMPI L, K, Pd, and Ma subscales were administered to 109 persons who were subsequently hired for the position of cottage attendant in an institution for the mentally retarded. 3 of the subscales and a composite score from all 4 of them distinguished significantly between attendants who were fired during the first 6 months of employment and attendants who were retained in employment and did not receive unfavorable evaluations from their immediate supervisors.

Although some studies (Kline, 1950; Yerburg, Holzberg, & Alessi, 1951) have found personality tests to be predictive of psychiatric aide or attendant performance, more methodologically sound investigations (Butterfield & Warren, 1962; Cuadra & Reed, 1957; Levine, 1951) have found personality tests to be of no help in the selection of attendants. Butterfield & Warren (1962) have suggested that one reason for the failure is that investigators have used no theoretical rationale in the selection of their personality measures. Rather, they have characteristically used a wide variety of measures in an attempt to find those which are, for some unspecified reason, predictive. The present investigation differs markedly from previous studies by using only a few personality measures each of which was expected to be predictive for some specified reason.

The present investigation also differs in another important way from a similar study by the same authors. In their earlier effort to predict attendant tenure, Butterfield & Warren (1962) used as predictors personality measures which had been secured over a period of 8 years. In the present study the predictor scores were collected in 18 months. The significance of this is that the supervisory staff in charge of hiring and firing attendants changed several times during the first investigation, but it remained relatively constant during the present study. An important source of error variance in the criterion scores, i.e., job tenure, which was present in the first study was not present in the present study.

METHOD

From January 1960 until July 1961, all applicants for the position of Cottage Attendant at the Oregon Fairview Home were given a group short form of the MMPI. This short form included the *L*, *K*, *Pd*, and *Ma* subscales. This test was not used as a basis for hiring attendants, however. Rather, the attendants were hired on the basis of a job application form and a short interview with one or two members of the institution's nursing staff. These individuals had had no training in interview techniques and little experience in hiring employees.

Three criteria were used by the interviewer in the evaluation of the applicants. If an applicant had listed several brief jobs in the work history section of the application form, he was not recommended for the job. No attempt was made to verify the work history reports of the applicants. If an applicant was grossly unusual in personal or physical appearance, he was not recommended for a job. For example, such things as slovenly or inappropriate dress, unshaven appearance, extreme obesity, etc., were counted against the person. Also, an attempt was made to evaluate the applicant's character by asking him if he drank. The interviewer was not interested in a "Yes" or "No" answer, but in the social poise which the applicant showed in responding to this presumably personal question.

During the 18-month period covered by this study, 109 persons were hired as cottage attendants. Forty-seven of these were discharged during the first 6 months of their employment and constituted the Discharged group for this study. Forty-eight attendants were employed for longer than 6 months and received no negative evaluations by their supervisors on a civil service rating form which was completed for each employee after 6 months of service. These 48 constituted the Still Employed group in this study. The other 14 employees who had been hired during the time this study covered were either discharged after more than 6 months of service or were retained as employees despite poor evaluations from their supervisors. The protocols of these attendants were not included in the analyses which follow.

PREDICTIONS

Five predictions were tested in this study. It was predicted that (a) dishonest persons (those with high MMPI *L* scale scores) and (b) defensive persons (those with high MMPI *K* scale scores) would have been able to hide poor work histories and undesirable characteristics from the hiring interviewer and that as these undesirable characteristics became apparent to their supervisors these persons would be discharged. It was also predicted that (c) persons with socially irresponsible tendencies (persons with high MMPI *Pd* scale scores) would be fired more frequently than more socially responsible people. It was expected, for example, that the high *Pd* would probably be irregular in his attendance, more likely to drink on the job, etc. It was also believed that (d) people who were very active and inclined to assume responsibilities which did not really belong to them (persons high on the *MA* scale of the MMPI) would have conflicts with their immediate supervisors and would consequently be more likely to be fired. Finally, it was predicted that (e) the greater the number of these characteristics which any attendant possessed, the more likely he was to be discharged.

RESULTS

Mann-Whitney *U* tests (Siegel, 1956) showed that the Discharged group had significantly higher *Ma* *T* scores ($p < .001$), significantly higher *Pd* *T* scores ($p < .01$), significantly higher *K* *T* scores ($p = .001$) and a nonsignificant trend ($p = .08$) toward higher *L* scale *T* scores than the Still Here group. Biserial correlations between the Discharged-Still Here dichotomy and *Ma*, *Pd*, *K*, and *L* scale *T* scores are .421, .400, .513, and .235, respectively. A chi square test (see Table 1) indicated ($p < .001$) that the greater the number of these individual signs which an attendant possessed the greater the likelihood of his being a member of the Discharged group. If an individual's *L*, *K*, *Pd*, and *Ma* subscale *T* scores were all less than 60, the chances were better than 2 to 1 that he would be retained, and if either three or four of these subscale *T* scores exceeded 60

TABLE 1
NUMBER OF *L*, *K*, *Pd*, AND/OR *Ma* SCALE *T* SCORES
GREATER THAN 60

Group	Number of <i>T</i> scores over 60			
	0	1	2	3 or 4
Still here	24	15	6	2
Discharged	10	12	15	11

Note.— $\chi^2 = 17.07$, $p < .001$.

the chances were better than 5 to 1 that the attendant would be discharged.

DISCUSSION

These results indicate that it is possible, given an understanding of the hiring and firing procedures of an institution, to predict, on a rational rather than strictly empirical basis, which attendants will probably be fired and which ones will probably be retained as employees. It is, perhaps, noteworthy that this is true even when the predictor variables are ones which were originally compiled by a strictly empirical procedure.

It should be noted that in the analyses reported 14 subjects were excluded because, although they had not actually been discharged, they had received unfavorable ratings. They, therefore, fell between the desirable retained employees and the undesirable discharged employees. This procedure probably inflated the relationship between the predictor and criterion variables.

It should be pointed out that although the retained-discharged criterion was used in this study, it is not the only one available and is probably one of the least desirable criteria from the standpoint of improving the quality of attendants at a given institution. It should be possible to secure objective measures of attendants who are the most highly regarded by their supervisors and to select for employment only similar applicants. It may also one day be possible to secure measures on attendants which are predictive of behavior changes in their charges. If this were true, then the ultimate criterion for hiring an attendant might be whether his personality and attitudes were such that he engendered desirable characteristics in patients for whom he cared.

REFERENCES

- BUTTERFIELD, E. C., & WARREN, SUE A. The use of the MMPI in the selection of hospital aides. *J. appl. Psychol.*, 1962, **46**, 34-40.
- CUADRA, C. A., & REED, C. F. Prediction of psychiatric aide performance. *J. appl. Psychol.*, 1957, **41**, 195-197.
- KLINE, N. S. Characteristics and screening of unsatisfactory psychiatric attendants and attendant applicants. *Amer. J. Psychiat.*, 1950, **106**, 573-586.
- LEVINE, S. The relationship between personality and efficiency in various hospital occupations. Unpublished doctoral dissertation, New York University, 1951.
- SIEGEL, S. *Nonparametric statistics for the behavioral sciences*. New York: McGraw-Hill, 1956.
- YERBERG, E., HOLZBERG, J., & ALESSI, S. Psychological tests in the selection and placement of psychiatric aides. *Amer. J. Psychiat.*, 1951, **180**, 91-97.

(Received April 11, 1962)

TEAM CREATIVITY AS A FUNCTION OF THE CREATIVITY OF THE MEMBERS¹

HARRY C. TRIANDIS, ALAN R. BASS, ROBERT B. EWEN, AND ELEANOR HALL MIKESELL

University of Illinois

The relationship between team performance and individual abilities was studied with creative tasks. Data from 3 experiments are presented which indicate that the conclusions of other investigators, who used manual dexterity and cognitive tasks, can be applied to creative tasks. Up to about 70% of the variance in dyadic creativity may be predicted from the individual creative abilities of the 2 members. Multiple correlations for the prediction of group performance from the knowledge of the abilities of the members did not improve when the interaction between the ability scores was considered. In 1 experiment, in which the procedure permitted the determination of the relative "dominance" of the 2 Ss, the correlations of the abilities of the dominant Ss with group performance were higher than the correlations of the abilities of nondominant Ss with group performance. The study is a 1st step towards the determination of the relationship of group and individual performance in groups composed of more than 2 individuals.

Several investigators have attempted to predict the performance of teams from the performance of their members. Comrey and his associates examined the relationship between team performance for dyads and the individual performances of the members of the team, using motor and cognitive tasks. Wiest, Porter, and Ghiselli (1961) have also studied this problem using dyads and a cognitive task. Two generalizations may be made from these investigations: (a) a fairly high proportion of the variance of team performance can be predicted from the knowledge of the performance of the members; (b) combinations of the individual scores, such as products, differences, sums, etc., and consideration of these new scores as separate variables does not appreciably raise the multiple *R* between individual performance and group performance. The present paper shows that these generalizations hold also in the case of team creativity.

Comrey (1953) had his subjects (Ss) work alternately on a manual dexterity task. He correlated the performances of the fastest

and slowest member and the team performance and obtained correlations of .52, .56, and .59 between the fast and slow member, the fast member and the team, and the slow member and the team, respectively. The multiple *R* after the *r*'s were corrected for attenuation was .66. This general pattern of findings was replicated by Comrey and Deskin (1954). When a cognitive task was employed (Comrey & Staats, 1955), the corresponding correlations were .25 (fast-slow), .76 (fast-team), and .67 (slow-team). In this case the multiple *R*, after the attenuation corrections, reached .91. Wiest, Porter, and Ghiselli (1961) obtained correlations of .28 (fast-slow), .72 (fast-team), and .63 (slow-team) with another cognitive task. The latter study's multiple *R* was .85.

The present paper indicates that similar results can be obtained with a variety of creative performances.

EXPERIMENT I

Method

Task A. This experiment was run during the summer of 1961, when President Kennedy's legislative program was beginning to take shape. The topic used involved legislation that was being debated in the press (peace corps, federal aid to education, school segregation, a depressed area bill, regulation of labor unions, medical care for the aged, the housing bill). Each problem was stated and contrasting points of view were presented arguing for and against some of the provisions of

¹ This work was made possible through Contract Number ONR 1834(36) with the Office of Naval Research on "Group and Organizational Factors Influencing Creativity" (F. E. Fiedler, L. Stolurow, and H. C. Triandis, Principal Investigators). We are grateful to F. E. Fiedler, W. A. T. Meuwese, and L. Stolurow for helpful critical readings of an earlier draft.

the bills. The student Ss, in groups of two (dyads), were instructed to discuss the issues and to produce a single inventive, creative, and original "solution" to each of these problems, in the form of suggestions for new legislation. The Ss were given half an hour for each problem. Each solution was mimeographed and judged by 30 judges (other students) on originality, practicality, and creativity. To obtain indices of creativity on an equal-interval scale, the judgments were handled by Thurstone's successive intervals procedure (Edwards, 1957). Thus, each solution acquired a scale value by the Thurstone method. The interjudge reliability of this procedure was .85, computed by the Hoyt (1941) method.

Task B. The Ss were given five concepts (e.g., Evolution, Army, Federal Aid to Education, Labor Unions, Immortality) and were asked to write an original, creative story containing all five concepts. Again, their creative products were mimeographed and administered to judges as described above. The interjudge reliability, for the creativity dimension, for this task was .90 (Hoyt, 1941).

Procedure. Forty-four 2-person teams were assembled for this experiment. Each S, working alone, wrote one "story" and solved one "problem" both before and after the team experiment. The measure of individual S's creativity was the average of his performance on these two occasions (scores correlated .61 and .50 for problems and stories, respectively). The team session required work on three problems (Task A) and three stories (Task B).

EXPERIMENT II

Forty-one dyads of male Ss worked on two problems: (a) "How can a person of average ability achieve fame and immortality though he does not possess any particular talent?" (b) "A church has completed about two-thirds of its building when it runs out of money. It is located in a blacklisted area in terms of credit. How can the church find the money to complete the building?"

Each S, and later each dyad, were asked to produce as many solutions as possible to these problems. The solutions were rated by two independent raters on a five-point scale of creativity. The raters were trained to give a large weight to the unusualness of the responses and a small weight to their practicality.

The exact procedure and the scores obtained from it were as follows: Each S was allowed 12 minutes to write as many solutions as he could to each problem and 3 minutes to choose his best solutions. This provided scores for the *actual* best solution given by the Ss (A_1 , A_2), for the *perceived* best solution (P_1 , P_2), and by adding up the ratings of all the solutions a *total* score (T_1 , T_2) was determined; the *number* of solutions (N_1 , N_2) and the *average quality* of the solutions (T_1/N_1 , T_2/N_2) were additional variables used in this analysis. After this the two Ss were allowed to work together and were instructed to produce solutions that did not overlap with their individual solutions. They were allowed 15 minutes to produce these solutions and 5 minutes to choose

the best of the solutions produced when working together as a dyad. This section of the procedure permitted us to obtain scores of dyadic actual best solution (A_d), perceived best solution (P_d), dyadic total score (T_d), number of solutions produced as a dyad (N_d), and *average* quality of solutions produced as a dyad (T_d/N_d). Following this the Ss were allowed 3 minutes to choose individually the best of all the solutions proposed (including their own, those of the partner, and those of the dyad). This provided scores for the *final perceived* best solution (FP_1 , FP_2). Finally, the Ss were given 7 minutes to discuss and to decide *together* which was the best of all the solutions proposed individually and by the dyad. This gave score FP_d —the *dyadic final perceived* best solution score.

The dominance of one of the Ss over the other can be ascertained from this procedure by looking at the pattern of their perceived best solutions. In most cases the FP score of one individual was identical to the FP_d score; in other words, one individual must have been able to convince the other that his particular judgment of which solution is the best should be adopted as the dyadic perceived best solution. This S was then judged to be the dominant one. For details of the interrelations between the 18 scores obtained from the procedure that is described above, and between dominance of the S and dyadic success, see Triandis, Mikesell, and Ewen (1962).

Procedure. Eighteen dyads worked on the Fame problem first and the Church problem second; 23 dyads worked in the reverse order. Thus, four replications of the relationships between the 18 variables mentioned above were obtained (two problems times two orders). In the present report we will discuss only those relationships which have a direct bearing on the determination of team creativity from the knowledge of the creativity of the members.

EXPERIMENT III

Thirty-three male and 32 female Ss participated. Only same-sex dyads were run. The creativity tasks were the same as those described in Experiment II. Further details of other aspects of this experiment can be found in Bass (unpublished).

In addition to the individual and dyadic creativity tasks, the Ss of this experiment had taken a battery of tests which included certain of the Guilford measures of divergent and convergent thinking (Guilford & Merrifield, 1960).² The divergent thinking measures were the Plot Titles, Consequences, and Alternate Uses tests. The convergent thinking measures included the Seeing Trends, Vocabulary Completion, and Object Synthesis tests. In addition, scores on the School and College Ability Test (SCAT) (Educational Testing Service, 1957) were available.

² We wish to express our appreciation to J. P. Guilford and the staff of the Aptitudes Research Project, University of Southern California, for permission to use the Plot Titles, Seeing Trends, Vocabulary Completion, and Object Synthesis Tests.

TABLE 1
CORRELATIONS BETWEEN INDIVIDUAL AND DYADIC PERFORMANCE
FOR PRESENT AND PREVIOUS STUDIES

Investigators and type of task	Correlation of fast (best) <i>S</i> and slow (other) <i>S</i>	Correlation of fast (best) <i>S</i> and dyad	Correlation of slow (other) <i>S</i> and dyad	Multiple <i>R</i>
Comrey (1953), manual dexterity	.52	.56	.59	.66 ^a
Comrey and Staats (1955), cognitive task	.25	.76	.67	.91 ^a
Wiest, Porter, and Ghiselli (1961), cognitive task	.28	.72	.63	.85
Experiment I				
Problems	.60	.69	.60	.73
Words	.40	.52	.60	.67
Experiment II				
Total (T)				
Fame	.70	.34	.47	.47
Church	.59	.35	.42	.47
Average (T/N)				
Fame	.63	.51	.62	.66
Church	.63	.48	.43	.52
Experiment III				
Males				
Total	.74	.84	.55	.85
Average	.82	.49	.34	.50
Females				
Total	.83	.58	.51	.58
Average	.72	.71	.78	.80

^a Corrected for attenuation.

RESULTS

Table 1 shows the correlations between the “best” and the “other” *S*s and the dyadic performance for both the previous studies mentioned above and our own three studies. This table indicates that the creativity tasks give somewhat higher correlations between the scores of the best and the other members than did the cognitive tasks of other investigators. The correlations obtained in the present study suggest that in our tasks the behavior of the *S*s becomes quite enmeshed. The multiple *R*s for the creativity tasks were *not* cor-

rected for attenuation. Had this been done the obtained values would have been similar to those of Comrey’s.

Table 2 shows the correlations obtained between various creativity scores of the dominant *S* and the dyad, from Experiment II.

Table 3 presents relationships between mental abilities and dyadic performance, from Experiment III, separately for male and female dyads. The table shows that the performance of dyads on the creativity tasks can be predicted from the abilities of the members, particularly the divergent thinking abilities.

The relationship between member's divergent thinking abilities and dyadic performance is particularly strong for the female dyads, permitting prediction of dyadic performance about as well as that obtained from individual performance on the same tasks. The relationship between the convergent thinking abilities and the creativity scores of the dyads is not as strong as for divergent thinking, except for the dyadic "average" score. This, however, is reasonable. Triandis, Mikesell, and Ewen (1962) have shown that T and N are scores reflecting the quantity of ideas (divergent thinking) while the average quality (T/N) is one of the purer quality scores. The relationships between the individual SCAT scores and dyadic performance appear relatively weak and for males these relationships

TABLE 2
CORRELATIONS BETWEEN MONADIC AND DYADIC PERFORMANCE FOR
DOMINANT AND NONDOMINANT SS IN EXPERIMENT II

Problem	Dominant S and non- dominant S	Dominant S and dyad	Nondominant S and dyad	Multiple R
Score: Actual (A)				
Fame (occurring first)	.43	.57	.46	.61
Fame (occurring second)	.04	.35	.12	.37
Church (first)	-.27	.32	-.52	.53
Church (second)	.26	.20	.10	.23
Score: Perceived (P)				
Fame (first)	.16	.27	.40	.45
Fame (second)	.08	.15	.25	.26
Church (first)	.10	.16	-.14	.21
Church (second)	.39	-.05	.25	.27
Score: FP ^a				
Fame (first)	.25	.71	.60	.83
Fame (second)	.01	.75	.15	.76
Church (first)	.02	.60	.19	.62
Church (second)	.57	.62	.59	.72
Score: Total (T)				
Fame (first)	.43	.49	.27	.50
Fame (second)	.24	.57	.28	.64
Church (first)	-.09	.50	-.15	.52
Church (second)	.41	.63	.14	.65
Score: Number (N)				
Fame (first)	.43	.28	.34	.36
Fame (second)	.38	.27	.33	.36
Church (first)	-.20	.41	.00	.42
Church (second)	.44	.77	.09	.82
Score: Average quality (T/N)				
Fame (first)	.33	.71	.61	.77
Fame (second)	.08	.52	.22	.56
Church (first)	.16	.57	-.01	.58
Church (second)	.55	.66	.58	.71

^a Predicting FP_d from the knowledge of the FP of the members.

TABLE 3
RELATIONSHIPS BETWEEN MEMBER ABILITIES AND DYADIC PERFORMANCE
(Experiment III)

Member abilities	Dyadic scores					
	T	N	T/N	T	N	T/N
	Males (<i>N</i> = 14 dyads)			Females (<i>N</i> = 16 dyads)		
Divergent thinking						
<i>X</i> ₁	.49*	.38	.43	.61**	.45*	.74***
<i>X</i> ₂	.29	.32	.18	.40	.31	.47*
Avg. of <i>X</i> ₁ and <i>X</i> ₂	.48*	.53*	.40	.56**	.41	.66***
Convergent thinking						
<i>X</i> ₁	.43	.54*	.09	.38	.30	.43*
<i>X</i> ₂	−.07	.18	−.22	.31	.29	.19
Avg. of <i>X</i> ₁ and <i>X</i> ₂	.19	.40	−.06	.30	.21	.39
SCAT						
<i>X</i> ₁	−.48*	−.35	−.30	.12	−.03	.41
<i>X</i> ₂	−.17	−.05	−.30	.31	.25	.45*
Avg. of <i>X</i> ₁ and <i>X</i> ₂	−.40	−.18	−.49*	.30	.16	.52**

Note.—*X*₁ is the higher of the two members of a dyad on the ability measure in question; *X*₂ is the lower member.
* *p* < .10, two-tailed test.
** *p* < .05, two-tailed test.
*** *p* < .01, two-tailed test.

are, in fact, negative. Divergent thinking correlated with the SCAT scores −.11 for males (*N* = 32) and .46 for females (*N* = 33); this difference was significant at the .05 level. Thus, it is not surprising that SCAT correlated negatively with dyadic performance for male Ss.

Improvement of multiple R by consideration of the X₁X₂ term. Previous writers have found that consideration of a term that is some function of *X*₁ and *X*₂ does not improve the prediction of dyadic performance. To check this possibility with our data we considered variable *X*₃ = *X*₁·*X*₂. Multiple *R*s

TABLE 4
CORRELATIONS BETWEEN MEMBER ABILITIES (*r*_{*X*₁*X*₂})
(Experiment III)

Member abilities	Males	Females
Divergent thinking	.80***	.56**
Convergent thinking	.37	.39
SCAT	.34	.55**

Note.—*X*₁ is the higher of the two members of a dyad on the ability measure in question; *X*₂ is the lower member.
** *p* < .05, two-tailed test.
*** *p* < .01, two-tailed test.

were computed for predicting dyadic creativity, from the knowledge of *X*₁, *X*₂, and *X*₃. In most cases this multiple *R* showed an improvement of the order of .01 over the multiple *R* obtained by the first two variables only. Thus, the results of previous investigators are supported by our own using the creativity tasks.

The lack of improvement in the multiple *R* is in part the result of the fact that a term, such as *X*₃, which is a function of *X*₁ and *X*₂, is of little value in increasing the multiple *R*. Any *X*₃ *must* be correlated with either *X*₁ or *X*₂, or both. Thus, Wiest et al. (1961) used *X*₃ = *X*₁ + *X*₂ and found a correlation of .605 between *X*₃ and *X*₁, and .936 between *X*₃ and *X*₂. They also used *X*₃ = *X*₁ − *X*₂ and obtained a correlation of .898 between this *X*₃ and *X*₂. This indicates that *X*₃ terms which are a function of *X*₁ and *X*₂ contain little variance that is not already present in *X*₁ and *X*₂ and, hence, are unlikely to account for much “additional” criterion variance. An additional consideration involves the selection of the best and other *S*’s performance as predictors of dyadic performance. The selection

of the best means that since $X_1 > X_2$, when X_2 is a small number, X_1 *can* be a small number, and when X_2 is a large number X_1 *must* be a large number. Thus, there is an artifactual correlation of X_1 and X_2 . According to Bereiter³ the expected value of this artifactual correlation in a bivariate normal population is .33. Now, given that X_1 correlates with X_2 at least to the extent of .33, this makes the addition of predicted variance by consideration of the X_1X_2 term much more difficult than it would have been if these two variables had a low or a negative correlation. Examination of Table 1 suggests that the other published studies also faced this problem—e.g., Comrey (1953) obtained correlation of .52 between X_1 and X_2 ; Wiest et al. (1961) obtained a significant correlation of .284 and that also was an artifact of the procedure. Thus, an approach such as the one we have used in our analysis of dominant and nondominant members of the dyads, may be more useful than the approach used so far in terms of the best and the other *S*. Comparison of the multiple *R*s obtained in Experiment II, in Tables 1 and 2, shows that those obtained with the dominant-nondominant *S*s are somewhat higher than those obtained by the best-other distinction.

The relative influence of the best and the other S. Comrey's results with dexterity tasks suggested that the performance of the dyad was more highly related to the performance of the *slow* than to the performance of the *fast S* (Table 1). With the cognitive tasks, however, this conclusion was reversed. Our results suggest that there is no particular consistency. Since the correlations between best and dyad on the one hand and other and dyad on the other hand are not significantly different from each other, and our results seem to be inconsistent, it is suggested that the influence of the two *S*s is about equal. The only set of consistent results is the set involving correlations between divergent thinking ability and our measures of creativity for males and females, in Tables 3 and 4. Here the best individual's scores are most closely related to the scores of the dyads than are the scores of the other individual. This latter finding is also consistent with the re-

sults of Comrey and Staats (1955) and Wiest et al. (1961) shown in Table 1.

The results of Table 2, on the other hand, suggest that for the quality scores established through our ratings (*A*, *T/N*) the dominant *S*s are consistently better predictors of dyadic performance than the nondominant. This conclusion, however, does not seem to hold quite as uniformly for the quantity scores (*T*, *N*). Nevertheless, the overall trend is quite clear: the dominant *S*s control performance much more than the nondominant.

DISCUSSION

Previous research has shown that from about 45 to 75% of the variance of team performance can be accounted for by the individual abilities of the team members. The tasks included both dexterity and cognitive abilities. It has also been shown that combinations of these individual scores do not increase the amount of variance accounted for by the individual scores. The present study is concerned with three experiments in which creative tasks were used. From about 22 to 72% of the variance of dyadic scores could be predicted from the knowledge of the individual creative abilities of the *S*s. Attempts to improve predictions by making use of the interaction between the abilities of the members ($X_1 \cdot X_2$) were unsuccessful. Thus, there seems to be fairly consistent support across tasks, investigators, and situations, for the proposition that substantial proportions of the variance of dyadic tasks can be predicted from the abilities of the members, but these predictions cannot be improved by consideration of interactions between the scores of the individuals.

It is an empirical question of some interest to establish the relationship between the amount of variance predicted and the number of *S*s who work in a team. It is likely that as the number of *S*s increases, the relative influence of some of the *S*s will become much greater than the influence of other *S*s. Thus, we can expect correlations between group performance and individual performance to become less similar for different team members than they were in our case, and to show significant differences. Already, inspection of

³ C. Bereiter, personal communication, 1962.

Table 2 shows that separation of the Ss into dominant and nondominant seems to have developed such differences between correlations. (In 19 out of 24 cases the dominant S's scores correlated more highly with the dyadic score than did the scores of the nondominant S.) Thus, the present study establishes rather firmly the generality of the relationship between group and individual abilities for dyads and can be the basis for comparisons of the relationships when the group size is larger.

REFERENCES

- COMREY, A. L. Group performance in a manual dexterity task. *J. appl. Psychol.*, 1953, 37, 207-210.
- COMREY, A. L., & DESKIN, G. Further results on group manual dexterity in men. *J. appl. Psychol.*, 1954, 38, 116-118.
- COMREY, A. L., & STAATS, CAROLYN K. Group performance in a cognitive task. *J. appl. Psychol.*, 1955, 39, 354-356.
- EDUCATIONAL TESTING SERVICE, Cooperative Test Division. *Cooperative school and college ability tests*. Princeton: ETS, 1957.
- EDWARDS, A. L. *Techniques of attitude scale construction*. New York: Appleton-Century-Crofts, 1957.
- GUILFORD, J. P., & MERRIFIELD, P. R. The structure of intellect model: Its uses and implications. *U. Sth. Calif. Psychol. Lab. Rep.*, 1960, No. 24.
- HOYT, C. The reliability obtained by analysis of variance. *Psychometrika*, 1941, 6, 153-160.
- TRIANDIS, H. C., MIKESELL, ELEANOR H., & EWEN, R. B. Some cognitive factors affecting group creativity. Technical Report No. 5, 1962, University of Illinois, Group Effectiveness Research Laboratory, ONR project 1834(36), Group and organizational factors influencing creativity. (Mimeo)
- WUEST, W. M., PORTER, L. W., & GHISELLI, E. E. Relationships between individual proficiency and team performance and efficiency. *J. appl. Psychol.*, 1961, 45, 435-440.

(Received May 10, 1962)

REWORDED VERSUS NEW INTEREST ITEMS¹

EDWARD K. STRONG, JR.

Vocational Interest Research, Stanford University

In revising the SVIB, 49 items were reworded and 101 items were replaced by new ones. The change in content between the original and reworded items was judged too slight to affect scoring of these items using current weights. The change in content between the original items and new items was judged too significant to allow these items to be scored on existing scales using current weights. To determine whether these judgments were correct, control and experimental groups responded to the items in a test-retest situation with a 3-day interval. 3 measures of stability agreed very well but not perfectly. An average of the 3 procedures indicated that 10 of the 49 reworded items should be classified as new instead of reworded. The criterion, stability of items, is a useful one in the selection of new items.

A measuring instrument which is used for a long period of time develops a certain amount of obsolescence. This has happened with the Strong Vocational Interest Blank (SVIB) since its last revision in 1938. Since its use has rested on the differences in responses of members of various criterion groups, and since the responses from these criterion groups have been obtained at very great expense, there is a serious loss when changes are made in items to which members of these groups have responded. Yet, after a considerable passage of time a number of the items in the Blank require change since they have either lost their meaning or have taken on new meanings.

In revising the item content of the SVIB to reflect changes in language usage, and in occupations, 101 items were replaced by new ones and 49 items were reworded (Strong, 1963, in press). A reworded item was defined as one in which the change was assumed to be too slight to cause appreciable differences in responses. New items were those in which the change was assumed to be so substantial as to tap a different source of response variance. Reworded items are to be considered available for continued scoring as though no change had been made. New items will not be used in any scoring scheme based on existing criterion groups. As time goes on the revised Blank will be employed in developing new scales and then the new items will be utilized.

¹ Ralph F. Berdie and David Campbell of the University of Minnesota provided the data discussed in this article.

Reworded and new items were so classified by the writer and several associates. The question which this report raises is whether the classification was accurate. To determine whether the classification was satisfactory two mimeographed Blanks were prepared, designated A and B. Blank A contained items as they appear in the present form of the SVIB. Blank B contained items as they would appear in the devised SVIB. There were 45 items identical to those in Blank A, 57 were reworded, and 12 were changed enough to be considered new items. Blank A was administered to a controlled group of 97 University of Minnesota male students on two occasions with an average of 3-day intervals between the two administrations. Blank A was given to an experimental group of 99 Minnesota students and Blank B 3 days later (see Table 1).

Test and retest of the Control group gave a measure of the stability of the 114 items. Test and retest of the Experimental group also gave a measure of the stability of the 45 unchanged items and, of primary concern here, a measure of how rewording the 57 reworded and 12 new items affected the responses.

How shall changes in responses to test and retest wording of items be measured? Seven procedures for measuring stability of items were given previously (Strong, 1943, ch. 25). Three of these procedures have been employed here. They are: percentage of identical responses, changes in attitude, and number of

TABLE 1
ITEM STABILITY: CONTROL GROUP VERSUS
EXPERIMENTAL GROUP

Test Blank A	L	I	D	Total
Control group, Retest Blank A				
L	7	1	0	8
I	0	77	3	80
D	1	6	2	9
Total	8	84	5	97 ^a
Experimental group, Retest Blank B				
L	1	3	0	4
I	22	39	23	84
D	4	3	4	11
Total	27	45	27	99 ^a

^a Expressing the total populations of 97 and 99 as 100 makes too slight a difference to warrant such additional calculations.

shifts to make the first response equal to the second.

METHOD

The three methods of calculating stability of items are illustrated in terms of Item 177, the original wording being System and the revised wording being Business methods magazines.

Calculation of number of shifts. For the Control group the shifts are $(8 - 8 = 0)$ between Test and Retest L, $(84 - 80 = 4)$ for I, and $(9 - 5 = 4)$ for D $\div 2 = 4$. For the Experimental group the shifts are $(27 + 4) + (84 - 45) + (27 - 11) \div 2 = 39$. (Number of shifts equals the largest of the three category differences, i.e., $84 - 80 = 4$ and $84 - 45 = 39$.)

Calculation of percentage of identical responses. For the Control group the percentage of identical responses equals $7 \text{ (LL)} + 77 \text{ (II)} + 2 \text{ (DD)}$ or 86. For the Experimental group the percentage equals $1 \text{ (LL)} + 39 \text{ (II)} + 4 \text{ (DD)} = 44$. Difference in identical responses = 42.

Calculation of differences in attitude. Attitude is here equivalent to the difference in liking and disliking. For the Control group the difference in attitude for test data is $8 - 9 = -1$, for retest data is $8 - 5 = 3$. The difference in attitude between test and retest is $-1 - 3 = -4$. For the Experimental group the difference in attitude for test data is $4 - 11 = -7$, for retest data is $27 - 27 = 0$. The difference in attitude between test and retest is $-7 - 0 = -7$. The difference between the two differences in attitude of the Control and Experimental group is $-4 - 7 = 3$.

The three procedures give comparable scores in most cases, but not in all. Usually an increase in liking is accompanied by a decrease in disliking, or the reverse. But occasionally there is an increase, or decrease, of both liking and disliking. When the former occasion occurs the number of shifts agrees with the change in attitude but when the latter occurs the number of shifts will be greater than the figure expressing change in attitude.

Unfortunately the problem is not the stability of responses to the two wordings of an item but whether the change in wording causes a change in the weighting. Weights assigned an item reflect the differences in category responses of a criterion group (C) and men-in-general (MIG). Whether rewording an item will cause a change in weights depends primarily upon the extent test and retest responses differ from MIG and only secondarily upon the extent they differ from each other, for which stability is a measure. Test and retest responses may differ appreciably and yet neither may differ from the responses of MIG so as to be weighted and again test and retest responses may differ very little and yet one may differ enough from MIG to be weighted and the other not. There seems to be no way to evaluate test and retest responses to an item in terms of changes in weights since relatively slight differences in test and retest responses yield significant changes in weighting and much larger differences cause no change in weighting. Consequently we must evaluate test and retest responses in terms of the amount of difference in responses between the two. In a very real sense we are concerned with the question, does the rewording of an item cause people to respond differently to the reworded item than they do to the original wording?

RESULTS

Number of Category Shifts

Table 2 gives the distributions of number of category shifts of the Control and Experimental groups. Where the items were identical in test and retest (first four columns of the table) the means ranged from 4.2 to 6.3 and the sigmas from 1.3 to 3.4. This value was

TABLE 2
SHIFTS IN RESPONSES: TEST VERSUS RETEST

Item	Control group, Test Blank A versus Retest Blank A			Experimental group, Test Blank A versus Retest Blank B		
	45 unchanged items	57 reworded items	12 new items	45 unchanged items	57 reworded items	12 new items
60						1
50						1
40						1
32					1	1
28					1	1
24						
20					2	2
16				1	3	1
12	1	1		1	3	
8	4	7		9	12	3
4	22	27	6	24	23	1
0	18	22	6	10	12	
Mean	5.1	5.1	4.2	6.3	8.5	27.2
Sigma	2.8	2.8	1.3	3.4	6.9	19.1

less than 16 for all but one of the 171 comparisons. If we assume that identical items may vary as much as 15 on this index over an interval of 3 days then we may also con-

clude that 7 of the reworded items having values in excess of 15 differ because of the change in wording. That is, they should be clasified as new not reworded items. The

TABLE 3
ITEMS RANKED ACCORDING TO THREE MEASURES OF STABILITY

Shifts in response		Percentage of identical responses		Changes in attitudes		Average of 3 measures		Revised classification
Rank	Item	Rank	Item	Rank	Item	Rank	Item	
1	177	1	177	1	384	1	374	New
2	384	2	333	2	50	2	177	New
3	333	3	156	3	333	3	333	Keep original wording
4	50	4	165	4	184	4	50	Keep original wording
5	222	5	50	5	222	5	156	New
6	184			6	328	6	184	Keep original wording
7	324			7	156	7	222	New
				8	80	8	80	Keep original wording
8	80	6	384	9	2	9	328	Keep original wording
9	156	7	80	10	151			
10	400	8	328					
		9	2					
11	328	11	400	11	165	10	165	New
13	151	12	151	13	177	11	2	
16	2	14	222	18	324	12	151	
18	165	17	184	40	400	13	324	
		18	324			14	400	

Note.—Rank 1 = lowest stability.

TABLE 4
WORDING OF ITEMS LISTED IN COLUMN 4 OF TABLE 3

Item	Original wording, Blank A	Revised wording, Blank B
384	Show firmness without being easy	Can be firm and show I mean it
177	System	Business methods magazines
333	Activity which produces tangible returns () () ()	Work for a definite amount of money () () ()
	Activity which is enjoyed for its own sake	Work which is enjoyed for its own sake
50	Landscape gardener	Landscape architect
156	Smokers	Stag parties
184	Social problem movies	Social problem movies or TV
222	Being pitted against another as in a political or athletic race	Competitive activities
80	Retailer	Store manager
328	Develop plans () () () Execute plans	Think up plans for doing something () () ()
		Carry out plans for doing something
165	Vaudeville	Variety shows
2	Advertiser	Advertising man
151	Drilling in a company	Drilling in a military company
324	Head waiter () () () Lighthouse tender	Head waiter () () () Lighthouse keeper
400	(1) Worry considerably about mistakes (2) Worry very little (3) Do not worry () () ()	Worry about mistakes A lot () Very little () Never ()

seven items are listed in Column 1 of Table 3, ranked in order of the number of shifts. Table 4 gives the wording of the items.

If our standard is lowered from 15 to 11, four unchanged items are classified as new items, together with 10 reworted items. The three additional reworted items that would be classified as new items are given in the second

section of Column 1 of Table 3 and the wording of the three items in Table 4 according to appropriate numbering of the items.

Percentage of Identical Responses

Table 5 gives the distribution of percentage of identical responses of the Control and Experimental groups. In the first four columns

TABLE 5
PERCENTAGE OF IDENTICAL RESPONSES: TEST VERSUS RETEST

Percentage of identical responses	Control group, Test Blank A versus Retest Blank A			Experimental group, Test Blank A versus Retest Blank B		
	45 unchanged items	57 reworted items	12 new items	45 unchanged items	57 reworted items	12 new items
90		1	1	1	1	
80	16	24	8	14	13	
70	24	27	3	26	28	
65	4	2		4	6	
60	1	3			4	
55					2	2
50					1	1
40					2	5
30						3
20						
10						
Mean	78.1	78.2	82.1	77.6	73.5	42.9
Sigma	5.7	6.8	6.2	6.2	10.1	12.2

there was no change in wording of the items between test and retest and the mean scores differed only slightly (77–82). If we assume that the distribution of 90–60 with the Control group is indicative of stability of items for the interval of 3 days then it may be assumed that percentage of identical responses of less than 60% is caused by changes in the wording of the test-retest items of the Experimental group. Lower agreement than 60% was recorded for five reworded items in the Experimental group. The five items that should presumably be classified on this basis as new not reworded are listed in Column 2 of Table 3.

If the standard is raised from 59 to 64 then four unchanged and four additional reworded items should be classified as new items. The four reworded items are also listed in Column 2 of Table 3.

Columns 5 and 6 of Table 5 show that only 5 of the 57 reworded items overlap with the 12 new items. Our classification of reworded and new items proves to be better than we believed would be the case.

Table 3 summarizes the data regarding shifts in responses (Table 2), percentage of identical responses (Table 5), and changes in attitudes (not recorded in detail). Items are ranked according to amount of difference in test and retest response of reworded items. Table 4 gives the original and revised wording of the items. Column 4 of Table 3 gives an average rank of the three methods of measuring stability. The poorest reworded item that clearly should be classified as new is Item 384. Many of our consultants objected to the original wording as ambiguous. It is a surprise to the writer that there should be a great difference in response to Item 177, System and the reworded Business methods magazines; and a still greater surprise that adding "or TV" to Item 184 should cause an appreciable change in response. There are many occasions when one must rely on judgment but this should never be done if there is any way to substantiate statistically one's judgment.

The upper third of Table 3 lists reworded items which are classified as new on the basis of what seems to be the best cutting point in the distributions. The middle group

of items are classified as new on a lower standard, as described above. The lower third of the table gives the ranking of items which are listed in three of the other four columns. The items are similarly grouped in Table 4.

RECOMMENDED CHANGES

The three measures of stability of items provide methods to determine whether a change in wording an item may be classified as inconsequential or not. The judgment of several experts was very good in their classification of changed wording as reworded or new. But 10 of the 57 changed items should have been classified as new not reworded. Eight of the 12 items classified as new were differentiated correctly by the shifts in response procedure and all were so differentiated by the percentage of identical response procedure.

By the time this experiment had been concluded a revision of the blank had been adopted. For obvious reasons it seemed desirable to make as few additional changes as possible. Consequently, it is recommended, see the last column of Table 3, that Items 384, 177, 156, 222, and 165 be classified as new not reworded; and that the original wording of the remaining five items be retained so that they will be classified as unchanged, not reworded. The data do not quite support the conclusion to keep the changed wording of Item 165 and classify it as new not reworded. Our consultants, however, all felt that the word Vaudeville was obsolete and the term Variety shows was preferable. For this reason the item is classified as new.

Previously it was stated that 101 items were replaced by new items and 49 items were reworded (Strong, 1963, in press). As a result of the above recommendations there are 102 new items and 43 reworded items. (Item 222 has been shifted from reworded to new and Items 50, 80, 184, 328, and 333 shifted from reworded to unchanged, original wording retained.)

DISCUSSION

Three measures of stability were employed agreeing very well in most cases. But the differences in attitudes procedure disagrees with the other two in a few cases and here it ap-

pears that the former is faulty. The data in Table 3 show slight change, for example, in difference of attitude of Item 177, System, and very large differences in measures of percentage of identical response and shifts in response. There is little basis for preferring one of these two latter procedures over the other but the shifts in response procedure more closely approximates the basic relations which affect weighting of items.

It seems reasonable that stability should be considered in selecting new interest items. Ambiguity of wording ought to be disclosed by measurement of stability. The data in this article support this conclusion.

It has been argued that items of one or only a few words are preferable to items of many words since a person could emphasize one word in an item of many words on one occasion and shift his emphasis to another word on a second occasion. There is, however, no relation between stability measured by percentage of identical responses and number of words per item. On the other hand, certain

groups of items are more stable than other groups. The more stable groups are Parts V, People; II, School Subjects; III, Amusements; and last half of Part VIII, Ratings of Characteristics. The less stable groups are Parts I, Occupations; IV, Activities; Part VII, Comparison between Two Items; and the first half of Part VIII, the poorest of all. In general the more stable parts of the Blank have fewer items than the less stable parts. It may be that the factor here is not number of words per item but the extent to which a person is habituated to the items. One is probably much more certain that he likes golf or algebra in contrast to being a carpenter or preferring to read a book rather than go to a movie.

REFERENCES

- STRONG, E. K., JR. *Vocational interests of men and women: 1943*. Stanford: Stanford Univer. Press, 1943.
- STRONG, E. K., JR. Good and poor interest items. *J. appl. Psychol.*, 1962, **46**, 269-275.

(Received May 14, 1962)

INFLUENCE OF DIGIT GROUPING ON MEMORY FOR TELEPHONE NUMBERS

FRANCIS T. SEVERIN AND MARILYN K. RIGBY

St. Louis University

4 distinct patterns of 60 7-digit numbers were compared for accuracy of dialing by 96 college students. The patterns consisted of the following number of digits separated by hyphens: 3-4, 1-3-3, 2-2-3, and 1-2-2-2. A memory drum was used for presentation of the digits and a telephone for dialing. The tapes for the memory drums contained 4 lists of the same digits, the lists differing in the pattern for each set of digits and in the sequence of the patterns. Data analyzed in a $4 \times 4 \times 4$ design for repeated measurement showed the 3-4 pattern, the pattern in current usage, to be superior to the others ($p < .01$).

Studies by Bell Laboratories (Karlín, 1958; Sinks, 1959) indicate that under controlled conditions all-numeral telephone dialing is somewhat faster and more accurate than the letter-numeral system it is gradually replacing. The finding that seven digits or two letters and five digits are learned with about equal ease is in agreement with memory span studies (Brener, 1940; Crannell & Parrish, 1957) reporting a slight advantage for digits over letters and unrelated words.

Seven-digit telephone numbers are printed in directories with a hyphen between the third and fourth digit. If the two clusters can be thought of as being related to quasi-memory spans, the introduction of additional hyphens to reduce the number of digits within a cluster should result in improved retention.

METHOD

Subjects

The subjects were 48 male and 48 female undergraduates enrolled in four sections of an introductory laboratory course at St. Louis University. Subjects were randomly assigned to four experimental setups consisting of identical rooms with duplicate apparatus. The experimenters were four paid laboratory assistants of advanced graduate standing each of whom tested 24 subjects.

A total of 60 telephone numbers was drawn at random from the Greater St. Louis directory to meet two restrictions after the exchange letters had been translated into the corresponding digits in the finger holes of the telephone dial: a digit was not to appear more than once in a given message, and no digits were to be arranged in the familiar order of counting.

Each telephone number was formed into four distinct patterns consisting of the following number of digits separated by hyphens: 3-4, 1-3-3, 2-2-3, and

1-2-2-2. Four lists of the 60 stimuli were constructed, each list containing the same digits in identical order, the stimuli differing only in pattern of hyphenation. For example, the number 5309642 appeared in the four lists as 530-9642, 5-309-642, 53-09-642, and 5-30-96-42. The arrangement of patterns within the lists was determined randomly with the restriction that no pattern appear in sequence more than three times, and that all patterns occur an equal number of times within each list. The same four lists were used in each of the experimental setups.

Apparatus

Four Lafayette-type memory drums were used to present the stimulus numbers. The four lists of messages were printed in elite type in four vertical columns on the paper tapes by an IBM tabulator which insured complete accuracy and uniformity. The order of the columns on the tapes was systematically varied for the four instruments and, to further minimize space errors, the windows used in the memory drums were rotated for successive subjects. Four new, Model 500 DR telephones were supplied through the courtesy of Southwestern Bell Telephone Company to test retention in a realistic manner.

Procedure

The subject was seated in a swivel chair facing a memory drum. He was instructed to study each number for the full 4-second period it appeared in the window, then to turn 90 degrees to his left, lift the telephone receiver and dial the number while saying the digits aloud. As a substitute for each digit he failed to recall he was to say the word "blank" and dial "operator." Verbalization of the digits enabled the experimenter to score for accuracy and also served as a partial control for rehearsing during the dialing operation. The memory drum was switched off at a blank space on the tape while the subject was occupied with the telephone. Four practice numbers were given to acquaint the subject with the procedure and the patterns to expect.

RESULTS

Each pattern of telephone message was scored both for the number of correctly dialed seven-digit messages and for the total number of incorrect and omitted digits. In order to isolate the variance attributable to the different experimental setups (experimenters) and the four arrangements in which the four different digit patterns were presented, the results were analyzed in a $4 \times 4 \times 4$ design with repeated measurements.

In the analysis of the correctly dialed numbers (see Table 1) the variance attributable to experimenter, pattern, and the interaction between pattern and arrangement were significant at less than the .01 level of significance. Similar findings were obtained from the analysis of the total number of incorrect and omitted digits (Table 2) at lower levels of significance for experimenter and for the interaction. Duncan's multiple-range test applied to the means of the four patterns in the first set of data showed the 3-4 pattern to be superior ($p < .01$) to the remaining three whose means did not differ from each other.

The significant difference between experimenters disappeared when the analysis was repeated without the data from Experimenter 4. An analysis of the latter alone showed a between patterns mean square significant at $< .05$ level of significance, and the interaction mean square just missed significance at that level. Setup 4, then, yielded the same form of

TABLE 2

ANALYSIS OF VARIANCE OF NUMBER INCORRECT AND OMITTED DIGITS			
Source	<i>df</i>	<i>MS</i>	<i>F</i>
Between (b) Ss	95		
Experimenters (<i>E</i>)	3	1333.10	2.23*
Arrangements (<i>A</i>)	3	204.71	
<i>E</i> × <i>A</i>	9	322.23	
Error (b)	80	412.13	
Within (w) Ss	288		
Patterns (<i>P</i>)	3	467.15	12.21**
<i>P</i> × <i>E</i>	9	26.39	
<i>P</i> × <i>A</i>	9	87.80	2.29*
Error (w)	267	38.25	

* $p < .05$.
** $p < .01$.

results reflected in the complete analysis in spite of the fact that all the scores were somewhat lower.

The interaction between pattern and arrangement is difficult to explain. An analysis of the random sequences of patterns by arrangement uncovered no anomaly which would seem to account for the details of the interaction. In the second, third, and fourth arrangements (see Figure 1), the 3-4 pattern

TABLE 1

ANALYSIS OF VARIANCE OF MESSAGES
WITH NO ERRORS

Source	<i>df</i>	<i>MS</i>	<i>F</i>
Between (b) Ss	95		
Experimenters (<i>E</i>)	3	261.19	6.44**
Arrangements (<i>A</i>)	3	17.41	
<i>E</i> × <i>A</i>	9	46.12	1.14
Error (b)	80	40.56	
Within (w) Ss	288		
Patterns (<i>P</i>)	3	43.90	16.02**
<i>P</i> × <i>E</i>	9	1.67	
<i>P</i> × <i>A</i>	9	11.49	4.19**
Error (w)	267	2.74	

** $p < .01$.

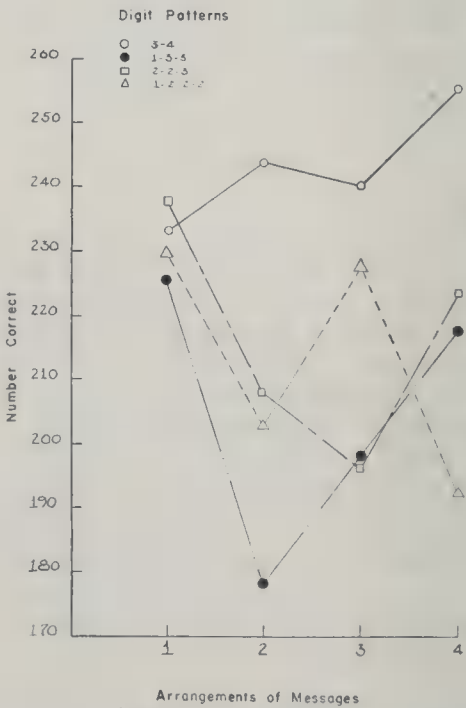


FIG. 1. Number of correct messages by pattern and arrangement.

was clearly superior to the others, whereas in the first arrangement all four tended to cluster together. The source of the significant interaction is apparently to be found in the other three patterns. Possibly irrelevant cues were more distracting to performance with the more difficult patterns.

DISCUSSION

The major finding was the superiority of the 3-4 pattern of digits under the experimental conditions described above. This is the pattern most similar to the familiar letter-numeral combination in current use in St. Louis. Since the experiment was deliberately structured as a telephone situation, it is likely that positive transfer effects from previous telephone dialing behavior were operative in the superior performance on the 3-4 pattern, just as practice in grouping influences memory span for digits and other materials (Crannell & Parrish, 1957; Martin & Fernberger, 1929; Oberly, 1928; Postman, 1954). The inferior performance on the other patterns may be related to negative transfer effects.

In a survey article, Conrad (1960) suggested that the optimum size of a group is three or four digits. The present findings are consistent with that generalization.

Two-thirds of the 3-4 pattern of numbers were dialed correctly. Under the conditions of

oral presentation of eight-digit messages to 24 telephone operators, Conrad (1958) found that 50% of the numbers were placed correctly when rehearsal was allowed, but only 35% when it was forbidden. In the present study rehearsal was allowed during the 4-second learning period although saying the digits aloud made rehearsal less likely during the dialing operation.

REFERENCES

- BRENER, R. An experimental investigation of memory span. *J. exp. Psychol.*, 1940, 26, 467-482.
- CONRAD, R. Accuracy of recall using keyset and dial: An effect of a prefix digit. *J. appl. Psychol.*, 1958, 42, 285-288.
- CONRAD, R. Experimental psychology in telecommunications. *Ergonomics*, 1960, 3, 289-295.
- CRANNELL, C. W., & PARRISH, J. M. A comparison of immediate memory span for digits, letters, and words. *J. Psychol.*, 1957, 44, 319-327.
- KARLIN, J. E. All-numeral dialing: Would users like it? *Bell Lab. Rec.*, 1958, 36, 284-288.
- MARTIN, P. R., & FERNBERGER, S. W. Improvement in memory span. *Amer. J. Psychol.*, 1929, 41, 91-94.
- OBERLY, H. S. A comparison of the spans of "attention" and memory. *Amer. J. Psychol.*, 1928, 40, 295-302.
- POSTMAN, L. Learned principles of organization in memory. *Psychol. Monogr.*, 1954, 68(3, Whole No. 374).
- SINKS, W. S. New numbers for tomorrow's telephones. *Bell Teleph. Mag.*, 1959, 38, 6-15.

(Received May 17, 1962)

EFFECT OF REHEARSAL OF TEMPORAL AND SPATIAL ASPECTS ON THE LONG-TERM RETENTION OF A PROCEDURAL SKILL¹

JAMES C. NAYLOR AND GEORGE E. BRIGGS

Ohio State University

The study examined several rehearsal techniques as means of facilitating the retention of a discrete procedural task. 4 rehearsal conditions were defined as: whole task rehearsal, temporal rehearsal, spatial rehearsal, and no rehearsal. All groups were trained for 5 days, given 10 days of no practice, 5 days of rehearsal, 11 more days of no practice, and a retention test. The number of commissive errors showed significant retention differences, with the whole rehearsal group performing best. Omissive errors and reaction time did not show group differences. It was also found that Ss emphasized those metrics of performance which gave the most immediate feedback.

In a recent review of skill retention over relatively long periods, Naylor and Briggs (1961) note that while there have been numerous studies of the effects of rehearsal on skill retention, (a) most of the research has utilized continuous skill tasks rather than those requiring discrete, procedural skills, and (b) no investigations have been reported which attempted to determine the influence of component rehearsal on retention of total task skills. The latter problem is of particular interest, since it is reasonable to expect that some component skills will be retained better than others in a complex task, and therefore differential rehearsal could be rather efficient, the more quickly forgotten skills receiving more rehearsal than those retained better over time from formal training.

Given a procedural task requiring a set of responses, there will be certain rules or restrictions placed on the appropriate sequence of these acts. For one thing, the sequence will involve an ordering of response occurrences in space, e.g., the checkout procedures on the flight deck of an aircraft require a spatially ordered sequence of button activations, knob turns, control deflections, etc., prior to flight. Further, a procedural task must be carried

out under certain time restrictions. In some tasks the time intervals between responses may be uniform, while in other cases they may be quite variable. Thus, all procedural tasks require responses ordered in both the temporal and the spatial domain, and the present study was concerned with the influence of rehearsal of these task requirements on the retention of a procedural task skill.

METHOD

Apparatus. A procedural task was employed which required the subject to learn a sequence of discrete responses in a particular spatial and temporal pattern. The display-response panel consisted of a column of nine pairs of lights with a red and an amber light in each pair, and a row of three response buttons associated with each pair of lights. The three response buttons were labeled OK, Emergency, and Check, respectively, and the subject was required to activate the appropriate button or buttons for each pair of lights, in turn, in order to "lock in" the amber lights. The process of locking in the nine amber lights had to follow a definite spatial and temporal sequence: spatially the subject had to respond to the lights in the order 1, 5, 2, 9, 8, 3, 6, 7, 4, where 1 represents the top pair of lights, 5 is the pair fifth from the top, etc., and temporally the interval between lights was 4, 8, 10, 4, 10, 6, 6, and 8 seconds, i.e., there was a 4-second interval between the onset of a light in the top position and the onset of a light in the fifth position from the top, there was an 8-second interval between onset of a light in the fifth position and onset of a light in the second position, etc.

It was possible for the experimenter to preprogram for each pair of lights one of three possible stimulus conditions: red light, amber light, or no light. The appropriate response or responses required of the subject depended upon which stimulus condition oc-

¹This research was carried out in the Laboratory of Aviation Psychology and was supported by the Aerospace Medical Division, Air Force Systems Command, under Contract No. AF 33(616)-7269 with Ohio State University. Permission is granted for reproduction, translation, publication, use, and disposal in whole or in part for any purpose of the United States Government.

TABLE 1
REQUIRED SEQUENCE OF RESPONSES FOR THE THREE POSSIBLE
STIMULUS CONDITIONS FOR EACH PAIR OF LIGHTS

Stimulus condition	Response(s) required ^a
1. Amber light glows at proper time	Press OK button to lock in amber light.
2. Red light glows at proper time	a. Press Emergency button to activate amber light b. Press OK button to lock in amber light
3. No light occurs at proper time	a. Press Check button to activate red or amber light b. If red, press Emergency button to activate amber light then press OK button to lock in amber light c. If amber, press OK button to lock in amber light

^a All responses had to occur within 3 seconds of the onset of the (proper) time interval; otherwise, a red light was locked in.

curred, and Table 1 lists the response contingencies for each possible condition. The particular stimulus conditions varied from trial to trial. However, the spatial and temporal order of the nine stimulus events remained constant throughout the entire experiment. Thus, the subject was required to learn (a) the invariant spatial order in which he was to react to the nine stimulus events (pairs of lights), (b) the invariant sequence of eight time intervals between each of the nine stimulus events, and (c) the required response or sequence of responses for each of the three possible stimulus conditions (red, amber, or no light).

If a particular stimulus event consisted of Stimulus Condition 1 (see Table 1) and the subject failed to activate the OK button within 3 seconds of the onset of that stimulus event, the red light was locked in automatically and he was scored with an omissive error. Also, if in Stimulus Condition 2 or 3 the subject failed to complete the required sequence of button activations within 3 seconds, the red light was locked in automatically and he was scored with an omissive error. If he pressed the wrong button for a particular stimulus event or if he pressed more than one button simultaneously for that event, he was scored with a commissive error or errors, depending on the number of "extra responses." However, in these cases he did not necessarily score an omissive error unless he failed subsequently to activate the appropriate buttons. Finally, in any case where the

subject activated a button *prior* to the onset of a stimulus event, he was scored with a commissive error for the extra response.

In addition to separate tallies for errors of omission and of commission, the experimenter recorded the reaction time for each trial (a sequence of nine stimulus events). Reaction time was measured as the interval between the onset of a stimulus event (amber light, red light, or no light) and the occurrence of the button response which locked in the amber light for that stimulus condition. A standard electric time clock, accurate to .01 second, accumulated the nine reaction times for each trial, and the experimenter recorded this total after each trial.

The subject received five trials on the first and on the last sessions and 10 trials during each of the intermediate sessions. During each trial, each of the three possible stimulus conditions (listed in Table 1) occurred an equal number of times. The experimenter developed 10 sets (trials) of nine stimulus events and this was a sufficient number so that the subject could not learn a particular sequence of stimulus events.

Experimental design and procedure. There were four groups of 17 subjects each and groups were defined by the characteristics of the rehearsal task (see Table 2). All subjects were undergraduate volunteers who were paid \$1.00 per experimental session. There were five daily sessions of training followed by a 25-day retention interval. The three experi-

TABLE 2
PRACTICE, REHEARSAL, AND RETENTION TEST SCHEDULES FOR ALL GROUPS

Group	Days 1-5 Initial training	Days 6-14 No practice	Days 15-19 Rehearsal	Days 20-29 No practice	Day 30 Retention test
1	Whole task	No practice	Temporal and spatial aspect	No practice	Whole task
2	Whole task	No practice	Temporal aspect	No practice	Whole task
3	Whole task	No practice	Spatial aspect	No practice	Whole task
4	Whole task	No practice	None	No practice	Whole task

mental groups (Groups 1, 2, and 3) experienced 5 days of rehearsal beginning 10 days after the last training session. The control group (Group 4) received no rehearsal. The rehearsal period was approximately midway through the retention interval. The subjects were assigned to groups on the basis of performance during the third training session.

The rehearsal conditions were as follows:

Group 1: These subjects experienced the same task as encountered during the training period, the whole task.

Group 2: These subjects rehearsed only the *spatial* aspects of the whole task previously practiced. All intervals between stimulus events were set at 7 seconds.

Group 3: These subjects rehearsed only the *temporal* aspects of the task. The spatial sequence of stimulus events was changed to 1, 2, 3, 4, 5, 6, 7, 8, 9. Thus, the events occurred in a simple top-to-bottom order rather than in the whole task order 1, 5, 2, 9, 8, 3, 6, 7, 4.

It may be seen, then, that Group 1 rehearsed *both* the temporal and the spatial aspects of the task, while Groups 2 and 3 rehearsed only the spatial or the temporal aspects, respectively.

RESULTS

The original acquisition, the rehearsal, and the retention test performance levels of the four groups are shown in Figures 1 and 2.

Figure 1 summarizes the omissive error data while Figure 2 provides the commissive error levels. Group averages were taken over five-trial blocks for each data point, except for the retention test session where individual trial data are shown.

Omissive errors. It may be noted from Figure 1 that group averages over the last few blocks of training trials indicate that subjects were approaching asymptotic performance prior to the retention period. During the first block of rehearsal trials Groups 2 and 3 show some deterioration in performance, which is probably due to the fact that this represented a “new” task, new in the sense that the subject was required to practice only one of the two aspects (temporal or spatial) of the procedural task. The subjects of Groups 2 and 3 did become quite proficient with the rehearsal task, as can be noted in Figure 1, during Rehearsal Blocks 3 through 9, when they generated substantially fewer omissive errors than did Group 1.

During the initial retention test trial all groups showed some forgetting, with Groups 2, 3, and 4 producing almost twice as many

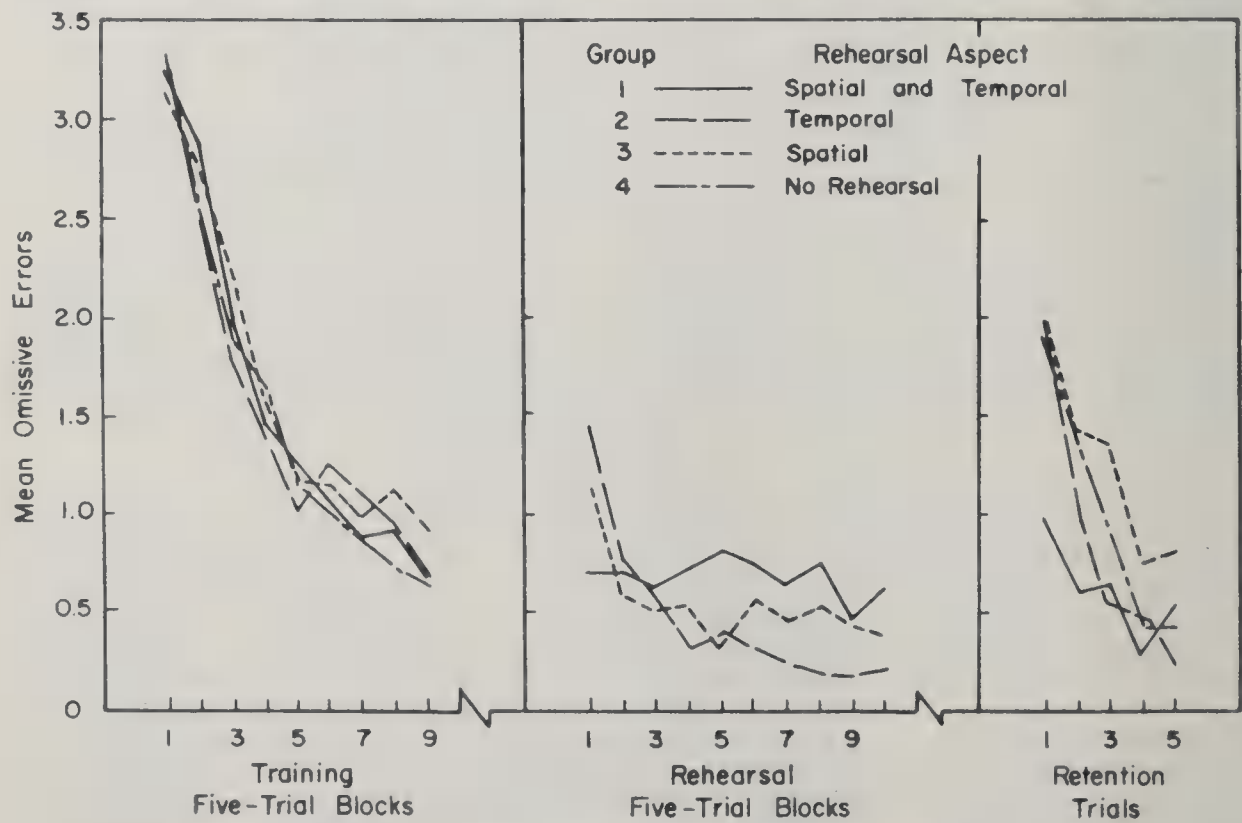


FIG. 1. Mean number of omissive errors for each group during training, rehearsal, and retest.

omissive errors as generated by Group 1. Recovery was quite rapid for all groups over the retention test trials; however, Group 3 appears to have lagged behind the other three groups in this regard.

An analysis of variance was performed jointly on the omissive error data of the five trials of the last training block and on the five trials of the retention test. The results of this analysis are summarized in Table 3. The test of rehearsal methods (Groups) indicated no significant differences among the four groups. However, since this analysis summed across both the training and retention blocks, separate analyses were run on each of the sets of data. Neither the last training block ($F = .746$, $df = 3/64$) nor the retention block ($F = 2.306$, $df = 3/64$) analysis showed differences between rehearsal methods. This was not unexpected, since as shown in Table 3 the Blocks \times Rehearsal interaction in the complete analysis also did not attain significance.

The significant Trials effect during the retention test block shown in Table 3 may be seen in Figure 1: there was a significant improvement in performance for all groups,

TABLE 3
ANALYSIS OF VARIANCE OF THE OMISSIVE ERROR SCORES
DURING THE LAST TRAINING SESSION AND THE
RETENTION TEST SESSION

Source	<i>df</i>	<i>MS</i>	<i>F</i>
Rehearsal methods (Groups) (G)	3	6.3	1.85
Subjects within Groups (<i>Ss</i> /G)	64	3.4	—
Blocks (B)	1	6.0	4.00*
Trials within Blocks (<i>Ts</i> /B)	8	9.1	15.17*
Trials within Training Block 9	4	0.8	0.75
Trials within retention block	4	17.5	29.17*
B \times G	3	2.0	1.33
B \times <i>Ss</i> /G	64	1.5	—
B \times <i>Ts</i> /B	24	1.0	1.00
<i>Ss</i> /G \times <i>Ts</i> /B	512	0.6	—
Total	679		

* $p < .05$.

even during the brief five-trial test session. The fact that the Trials effect was not significant during the last training block would indicate that the groups had approached closely an asymptote in performance at the end of training. It should be noted also that the significant Blocks effect does indicate that the groups performed less proficiently during

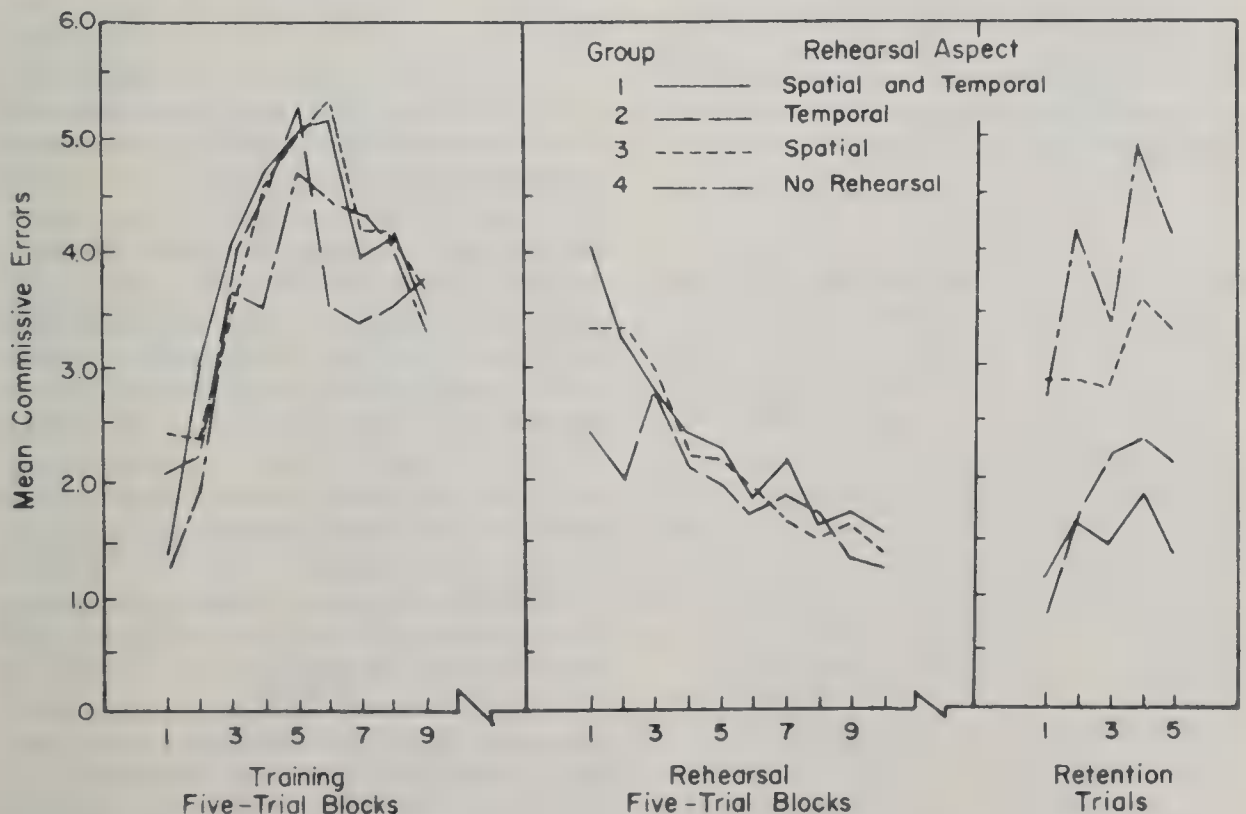


FIG. 2. Mean number of commissive errors for each group during training, rehearsal, and retest.

retention test than at the end of training. However, as has been mentioned, this loss in skill was apparently equivalent for all groups.

Commissive errors. Figure 2 and Table 4 show, respectively, the performance curves and the related analysis of variance for the commissive error data. As was the case with omissive errors, the overall analysis showed no effect due to rehearsal schedules. This was *not* substantiated, however, by separate analyses on the training and on the retention blocks. While group differences during the training block were not significant ($F = .088$, $df = 3/64$), the differences during retest *were* significant ($F = 6.838$, $df = 3/64$) at $p < .01$. Thus, in the case of commissive errors the rehearsal techniques did differentially affect performance during the retention test trials (see Figure 2). The best performance was attained by Group 1 which rehearsed on the entire task, while Group 2 (temporal rehearsal) performed nearly as well. In fact, Group 2 actually was superior to Group 1 on the initial retention test trial in terms of fewer commissive errors.

Groups 3 and 4 both exhibited marked losses in retention as compared to Groups 1 and 2, with Group 4 (no rehearsal) showing a somewhat greater loss than Group 3 (spatial rehearsal). This difference is supported also by the significant Blocks \times Rehearsal interaction which indicates that Groups 3 and 4

TABLE 4

ANALYSIS OF VARIANCE FOR THE COMMISSIVE ERROR SCORES DURING THE LAST TRAINING SESSION AND THE RETENTION TEST SESSION

Source	df	MS	F
Rehearsal methods (Groups) (G)	3	47.7	1.76
Subjects within Groups (Ss/G)	64	27.1	
Blocks (B)	1	160.0	16.16*
Trials within Blocks (Ts/B)	8	13.2	6.60*
Trials within Training Block 9	4	13.0	6.50*
Trials within retention block	4	13.5	6.80*
B \times G	3	55.0	5.60*
B \times Ss/G	64	9.9	
G \times Ts/B	24	2.2	1.10
Ss/G \times Ts/B	512	2.0	
Total	679		

* $p < .05$.

TABLE 5

PRODUCT-MOMENT CORRELATIONS BETWEEN MEAN COMMISSIVE ERRORS AND MEAN OMISSIVE ERRORS FOR EACH OF THE FOUR GROUPS DURING EARLY TRAINING (Trials 1-25), LATER TRAINING (Trials 26-45), AND DURING RETENTION TEST

Group	Trials 1-25	Trials 26-45	Retest
1	-.94**	.29	-.87
2	-.87**	-.18	-.97*
3	-.86**	-.19	-.83
4	-.93**	-.22	-.78

* $p < .05$.
** $p < .01$.

show greater relative losses in skill than Groups 1 and 2.

The significant Blocks effect of Table 4 may be seen in Figure 2: performance for all groups was better at the end of training than during retention test. Also of interest was the fact that Trials effect was found to be significant in *both* the training and retention blocks. Apparently, then, the groups had not reached an asymptotic level of performance with regard to commissive errors at the end of training, and the decrements in skill suffered over the retention interval were rapidly overcome by practice during the retention test.

Reaction time. These data were completely insensitive to rehearsal effects, i.e., there were no significant differences among groups either during the rehearsal period itself or during the retention test trials. Therefore, no presentation of the data is warranted. It was noted that average reaction time was quite asymptotic from the fifth through the ninth (last) training block, and it would appear that this aspect of skill was rather highly overlearned during original training. It follows, then, that any possible differential effects of the rehearsal variable would not be reflected in these data.

Correlational analysis. Due to the design of the apparatus, it was possible for the subject to reduce or minimize the number of omissive errors (red lights) on a given trial by making a number of commissive responses. For example, if a red light appeared as a stimulus and the subject pressed the OK button first, this would not be a correct response.

However, if he were able to respond quickly enough, he then could press the Emergency button followed by the OK button (the correct response pattern) and avoid making an omissive error. His total number of responses to that light would be three, one more than the minimum number necessary to avoid a red light.

Because of the compensating arrangement between these two metrics, it was of interest to determine the actual relationship of commissive and omissive errors for the four groups. Table 5 shows the product-moment correlations between the measures during training and during retention test for each group. During the first 25 trials the correlations were all highly negative, in all cases being statistically significant. These relationships may be seen by comparing Figures 1 and 2 (Blocks 1-5). However, the correlations between measures during the last 20 trials of training (Blocks 6-9) were all non-significant, with Group 1 showing a slight positive relationship and all others being slightly negative. During retest, the measures were again highly negative, although the only group demonstrating a relationship of $p < .05$ was Group 2.

DISCUSSION

Only one of the three metrics of performance—that of commissive errors—demonstrated clearly the utility of rehearsal as a means of maintaining or facilitating the retention of a discrete procedural task. In all cases, the differences were in the expected direction, but only with the commissive errors were the differences large enough to attain statistical significance. One possible explanation for this lies in the finding that only the commissive error measure failed to indicate asymptotic performance at the end of training. It may be seen in Figures 1 and 2 that omissive error performance had leveled off by the fifth day, while subjects were still improving with regard to commissive responses. Thus, the sensitivity (and utility) of rehearsal may well be a function of the degree to which the response characteristics have been learned or established, being less effective with those skill components which have been more thoroughly acquired.

Examining the results obtained with commissive errors, one finds that the whole-task rehearsal procedure was markedly superior, a result which was expected. Rehearsal on only the temporal aspect of the task also led to increased skill during retention test, although not to as great an extent as whole practice. Spatial practice, on the other hand, was virtually no better than no rehearsal at all.

The overall superiority of whole-task practice can probably be attributed to two factors. First, and most obvious, in the case of single-dimension rehearsal (Groups 2 and 3) that dimension not being rehearsed will suffer a loss even though skill in the other dimension is maintained. Second, with a procedural task, both dimensions (temporal and spatial) are always present to some degree. Thus, in this experiment, while it was possible to *emphasize* the spatial rehearsal dimension by making the timing less random, it is impossible to devise the task without a time dimension. Similarly, it is impossible to design a procedural task without a spatial dimension. Therefore, it may be that practice on the less complex dimension during rehearsal actually interferes with retention of the more complex dimension present in the whole task.

The fact that rehearsal on the temporal dimension was more efficient than spatial rehearsal would indicate that the time dimension is inherently more difficult than the spatial, and that perhaps timing is a more critical aspect of retention for discrete tasks than is order of responding. The fact that practice on the apparently more difficult (temporal) dimension resulted in greater whole-task performance as opposed to practice on the easier dimension supports the general notion of Bilodeau (1955) and Bilodeau and Bilodeau (1954). They found that practice on separate task dimensions was increasingly beneficial as the difficulty of the dimension was increased. While they did not use retention as a criterion, a rational extrapolation of this notion to rehearsal and retention conditions seems appropriate.

The concept of interference or inhibition as a function of rehearsal on a less complex dimension is supported by the studies of Shepard (1950) and Frankmann (1957). They report evidence that conflicting practice

may deter retention over short periods of time. However, the classical explanation of the amount of such inhibition being a function of task (dimension) similarity does not explain the relative efficiency of temporal and spatial rehearsal, only that both should be less efficient than the whole method. With temporal rehearsal, the spatial dimension was reduced (in terms of information) from 2.8084 to 0.0000 bits. Similarly, with spatial rehearsal, the temporal dimension was reduced from 2.8084 to 0.0000 bits. Thus, the degree of relative similarity of the nonemphasized dimension to the dimension present in the retention task was the same for both rehearsal conditions. This would suggest that the primary difference between rehearsal procedures was not due to inhibition but to the basic, intrinsic difficulty differences of the temporal and spatial dimensions.

The high negative relationship found between the omissive error metric and the commissive error metric early in practice indicates that subjects were more concerned with omissive errors. This was not unexpected since an omissive error provided the most immediate feedback in the form of a red light at the end of each stimulus event. Thus, during the first 25 trials, the subject was able to reduce the number of omissive errors drastically by making a number of extra responses (commissive errors)—thus, the high negative relationship noted in Table 5. By the end of Trial 25, most subjects had reached a fairly high level of proficiency as measured by omissive errors, and they were able to concentrate upon reducing the number of extra responses during the remaining four training blocks—thus, the generally low and nonsignificant relationships

found in Table 5 for Trials 26–45. These correlations were probably reduced in size due to the fact that subjects were approaching asymptote on the omissive error metric. It is interesting to note that in the retention test situation the subject again reverted to the initial behavior pattern by using greater numbers of commissive errors to reduce the number of omissive errors.

The data of Table 5 provide, therefore, a rather interesting insight to the method used by the subject in accomplishing skill at this procedural task. It is apparent that he was able to reduce the number of omissive errors by committing a number of extra responses (commissive errors) and that he was particularly prone to accept this expedient during those periods when he was uncertain of the spatial and temporal aspects of the task (during the initial training and during the final retention test trials).

REFERENCES

- BILODEAU, E. A. Variations in knowledge of component performance and its effects upon part-part and part-whole relations. *J. exp. Psychol.*, 1955, **50**, 215–224.
- BILODEAU, E. A., & BILODEAU, INA. The contributions of component activities to the total psychomotor task. *J. exp. Psychol.*, 1954, **47**, 37–46.
- FRANKMANN, JUDITH P. Effect of amount of interpolated learning and time interval before test on retention in rats. *J. exp. Psychol.*, 1957, **54**, 462–466.
- NAYLOR, J. C., & BRIGGS, G. E. Long-term retention of learned skills: A review of the literature. *USAF ASD tech. Rep.*, 1961, No. 61-390.
- SHEPARD, A. Losses of skill in performing the standard mashburn task arising from different levels of learning on the reversed task. *USN Spec. Dev. Cent. tech. Rep.*, 1950, No. 938-1-9.

(Received May 22, 1962)

CHANCE ON SVIB:

DICE OR MEN?

DAVID CAMPBELL

University of Minnesota

The profile report form for the Strong Vocational Interest Blank has shaded areas, established by throwing dice, to indicate chance scores on the various scales. The SVIB manual suggests that this shaded area be used as a reference point to determine if a given score is significant. This article suggests that the shaded areas should be determined by scores of the Men-in-General group instead of throwing dice. The effect of this on the shaded area is shown on a sample profile report form.

On the profile form used to indicate the individual's results on the Strong Vocational Interest Blank (SVIB), the shaded area on each scale is the "chance" level. Strong reports (1959) that these areas were determined by throwing dice, and:

Significant scores may be defined as those which are four standard score points above or below scores likely to occur by chance. Scores which fall in the shaded areas may be interpreted as easily obtainable by chance and therefore indeterminate for the particular scale in question (p. 7).

Four standard score points are used because that is roughly the standard deviation of the distribution of chance scores.

Stephenson's recent research (1961) on scores in the chance range has provided the stimulus for further thought on this subject. He has shown that scores in the chance range are just as stable over a 10-year test-retest period as scores any place else in the range, and are not random scores in the sense that tosses of the dice are random.

Perhaps a better way of deciding whether or not a score is significant of interest in a given occupation would be to compare that score with the distribution of scores obtained by Men-in-General (MIG), not with the distribution of scores obtained by tossing dice. The purpose of the SVIB is to indicate explicitly how similar the individual's interests are to any particular Criterion group, and, at least implicitly, how different these interests are from MIG. An explicit comparison of an individual's score with the MIG distribution would seem to be very helpful in interpreting scores.

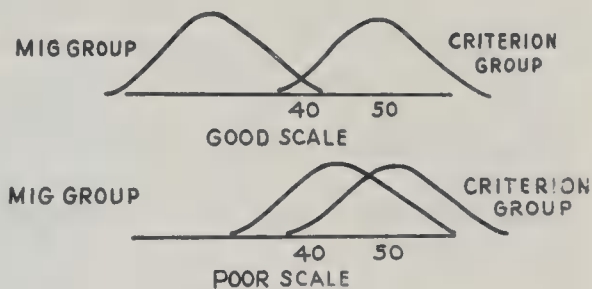


FIG. 1.

It would also allow users of the SVIB to take into account differential validity of the various scales. For example, if we take two hypothetical scales, one that separates the Criterion group very well from the MIG and one that does a much poorer job, the overlapping distributions might be as given in Figure 1.

If an individual receives a standard score of 40 on both scales, the current methods of interpretation would consider these two scores equally indicative, particularly if the chance areas were reasonably similar. However it is apparent that considering the range of scores on each scale in the MIG group, one score is considerably more significant than the other.

Following this reasoning, it would seem more reasonable, on the SVIB profile form, to shade the sigma range (mean ± 1 SD) for the MIG distribution rather than the chance dis-

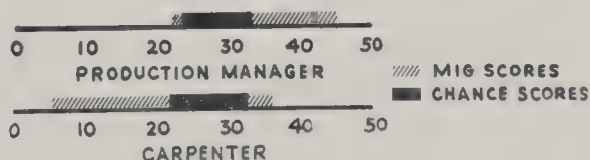


FIG. 2.

STRONG VOCATIONAL INTEREST TEST-MEN

HANKS REPORT FORM FOR-

SEE OTHER SIDE FOR EXPLANATION

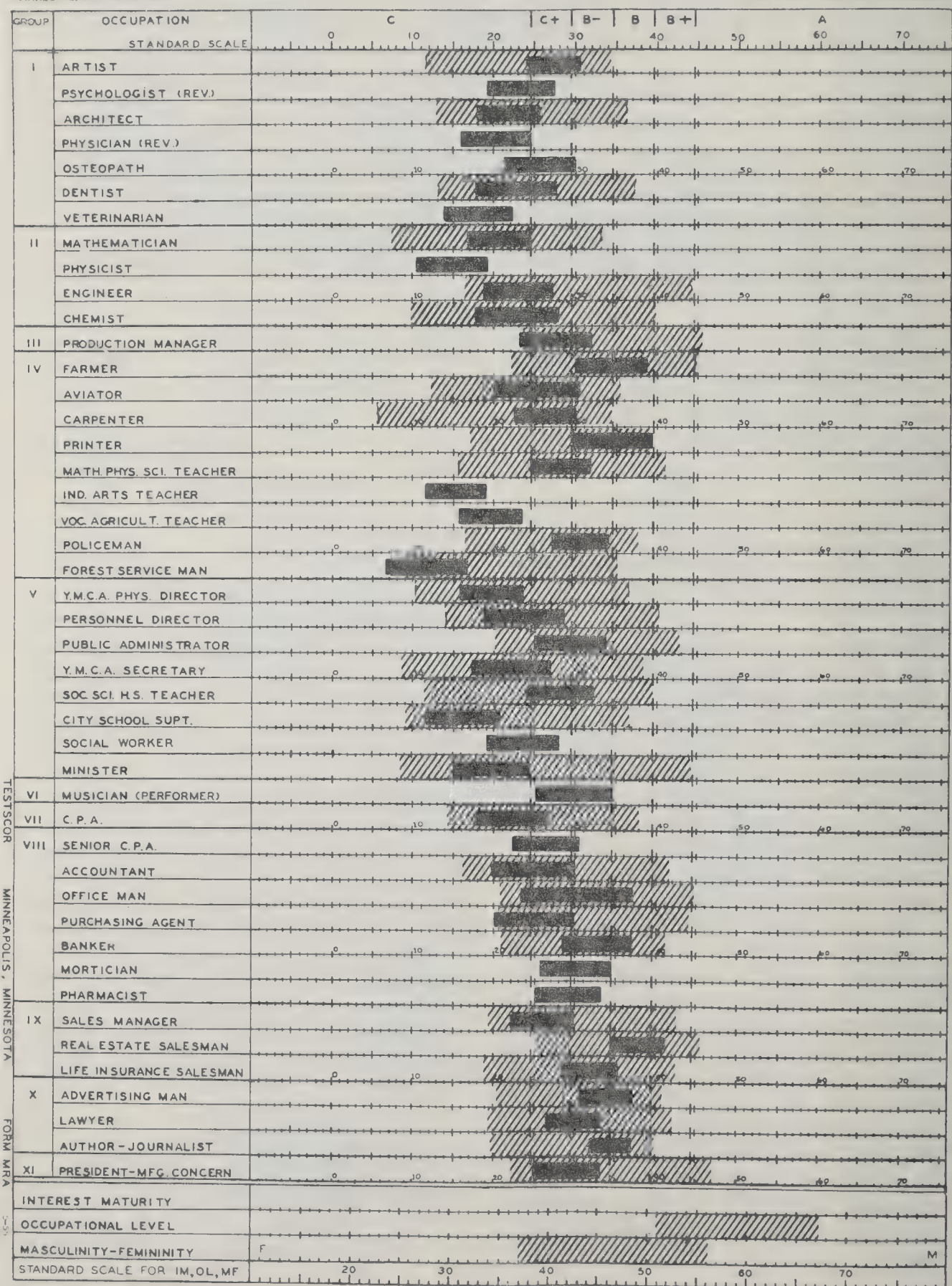


FIG. 3

tribution. (Or of course both could be indicated in some way.)

The following example shows how this might be done. Two scales, Carpenter and Production Manager, have very similar chance areas, but different distributions of MIG scores. Figure 2 shows this difference graphically.

If an individual scores 40 on both scales, most users give equal weight to these scores. Both scores are about equidistant from the chance area, and they both fall at the same point when compared to the occupational group. But when they are compared with the MIG distribution on the respective scales, the interpretation is changed. A standard score of 40 on the Production Manager scale falls well within the MIG group, while on the Carpenter scale only a few percent of the MIG reach a score of 40. Thus on the Carpenter scale a score of 40 would seem to be considerably more meaningful than on the Production Manager scale.

Considerable support from this viewpoint comes from asking counselors how they treat the shaded area. A sizable number of the counselors on one counseling bureau staff comment something to the effect that, "Yes, I know those scores were determined by throwing dice, but I usually think of them as representing man-in-general in some manner."

Figure 3 shows how the complete profile form would look if both chance and MIG sigma ranges were indicated. The solid area is the usual chance shaded area, the cross-hatched area is the MIG. All data were taken from the SVIB manual.

REFERENCES

- STEPHENSON, R. R. Chance versus nonchance scores on the SVIB. *J. appl. Psychol.*, 1961, **45**, 415-420.
STRONG, E. K., JR. *Strong Vocational Interest Blank manual*. Palo Alto: Consulting Psychologists Press, 1959.

(Received May 22, 1962)

VIGILANCE PERFORMANCE UNDER CONDITIONS OF REDUNDANT AND NONREDUNDANT SIGNAL PRESENTATION

WILLIAM C. OSBORN, RICHARD W. SHELDON, AND ROBERT A. BAKER

United States Army Armor Human Research Unit, Fort Knox, Kentucky

Brief interruptions in a sound, a light, or both the sound and light, were monitored by 41 Ss over a 3-hour period. A dual response apparatus allowed the Ss to report the light signals, the sound signals, or both. The detection rate was found to be significantly better for the redundant signals than for either alone. While the detectability of each component of the redundant signal was comparable to its corresponding single mode, a systematic deviation in the bimodal curve—from predicted to observed—was noted. It was concluded that the weaker component of a redundant signal contributes significantly to the overall detectability, and the use of dual channel displays in applied vigilance situations is justified.

In attempting to reduce the performance decrement typically found in vigilance tasks, several experimenters have presented the signal simultaneously to the eyes and the ears. Such "redundancy," it is argued, should lead to an improvement in the rate of detection. On the theoretical side, while such factors as differential stimulus intensity, intersensory interaction, and individual differences in the preference of one sensory channel over another may be of considerable significance, the principal issue is one of sensory summation: Does or does not bimodal sensory data summate to a level greater than that of the better single modality?

Initial findings showing the general superiority of combined audio-visual presentation (Schafer & Shewmaker, 1953) have been supported in more recent studies involving detection tasks. Loveless (1957), comparing single and combined audio-visual signals in short watch sessions, found the combined signals produced a higher detection rate. Using auditory and visual displays, individually and in combination, Buckner and McGrath (1961) confirmed the increased detection probability for the redundant presentation. Baker, Ware, & Sipowicz (in press), in a partial replication of the Buckner study, reported a similar summation effect over a longer watch. They failed, however, to obtain significance between the redundant condition and the easier (auditory) of the two modes. They also suggested that individual differences in ability to use a single

versus a dual channel display may determine the outcome.

In estimating the amount of summation under the redundant condition, both the Loveless and Buckner and McGrath studies found that predictions from the separate rates of detection consistently overestimated the redundant curve. Yet, in both cases the investigator based his predicted curve on group data which, as Buckner and McGrath noted, fails to control for the inflating factor of positively correlated auditory and visual performance.

Thus, in spite of the considerable evidence supporting a summation hypothesis for bimodal data, neither the extent of this effect nor the specific roles of the contributing factors are clear. Whether data from the component modalities summate strictly in an informational sense, or whether it is modified either by intersensory masking or by a facilitatory effect, remains unanswered. Moreover, the possibility that the higher bimodal function is simply the result of the cumulative detections of several monitors' preferred or "stronger" sensory channels has yet to be rejected. Accordingly, the present study was designed with the specific intent of analyzing the summation effects over an extended vigilance session. In this study separate response channels for the auditory and visual modalities were provided as a means for examining the relative contribution of the two compo-

nents to the total detectability of the redundant signal.

METHOD

Subjects. The subjects (Ss) were 41 enlisted trainees with no apparent visual or auditory defects as determined by Army medical standards. The Ss' average age was 20.2 years.

Apparatus. The sound level base for the auditory signal was produced by a Grayson-Stadler noise generator, Model 455B, operating at 36.5 db. SPL and presented to S over PDR-8 earphones. The visual signal occurred as an interruption in the output of two adjacent General Electric 12-16 v. bulbs, No. TS 53, operating at $9\frac{1}{2}$ v. ac mounted 3 inches back from a $\frac{3}{4}$ -inch opening in a flat black plywood box at approximately eye level. The bulbs, fixed with red and green lucite lens caps, produced hues which, when passed through a double thickness of standard tracing cloth (Imperial No. 125B), appeared as a brownish light. Fixed at the base of each display were two response buttons, separately marked "sound" and "light." Using a randomized Mackworth schedule (1950), of 24 signals per hour, signal interruptions of .041-second duration (measured at the timing relay with a Hunter Klockounter) were produced by a Gerbrands variable interval programmer and a simple timing circuit. Using a table of random numbers, the intersignal intervals were also varied to eliminate the likelihood of Ss developing any useful "expectancies." Serial occurrence of the three signal types, auditory, visual (nonredundant), and combined audio-visual (redundant), was also randomized within blocks of 30 minutes (four of each signal type), and was controlled through a programed stepping switch. Signal presentations and Ss' signal detections were recorded on a 20-pen, Model AW, voltage type, Esterline-Angus operations recorder.

Procedure. The Ss' task was to detect aperiodic interruptions of continuous light, sound, or light and sound sources during a 3-hour watch session. Eight isolated rooms on the second floor of a temporary Army barracks were used for monitoring. All windows were painted flat black, and room illumination was furnished by a single 60-watt overhead bulb. A chair and table with the monitoring display were situated along one wall of the room, and a circulating fan producing approximately 46 db. SPL ambient noise was located along the opposite wall.

After the Ss were acquainted with the general purpose of vigilance research, they surrendered their watches, pens, reading material, etc., and were assigned to individual rooms. Instructions were then given over the headsets:

This morning you have two vigilance tasks to perform at the same time. One is to monitor the light in the center of the box in front of you; the other is to monitor the steady noise which will be coming over your headset. The signal you are to watch for will be a very short interruption in either the light or the noise, or in both the light

and the noise at the same time. When you see the light blink off you are to press the button on your left marked "light"; when you hear a break in the noise, press the button marked "sound"; and when both the light *and* the noise cut off at the same time, you are to press *both* buttons. The signals will be very brief, so you should attend closely and try to pick up each one.

A short practice period consisting of 10 signals of each type, spaced 5 seconds apart, was then given. Upon completion of the practice session, Ss were informed:

From here on you may expect any one of the three types of signals at any time.

RESULTS

In the present analysis a "detection" was scored on a redundant signal if at least one component (either auditory, visual, or both) was reported. The mean percentage of detections, analyzed by 30-minute periods, for each signal condition is shown in Figure 1. On an overall basis, detection was highest for the redundant (R_o) condition, next highest for the auditory signal (NR_a), and lowest for the visual (NR_v) signal. A variance analysis of the number of detections (Table 1) showed a significant difference among the signal conditions. It should be noted, however, that the visual curve was slightly higher than

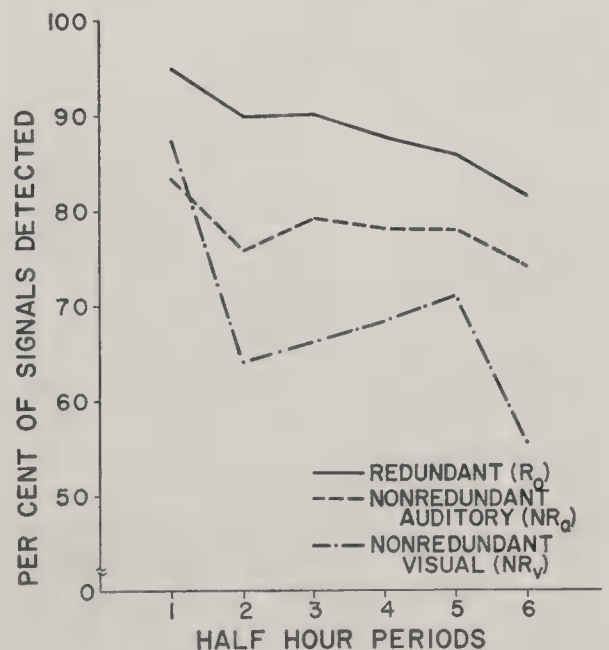


FIG. 1. Performance on the nonredundant auditory, nonredundant visual, and redundant signal types over the 3-hour vigilance session.

TABLE 1
ANALYSIS OF VARIANCE OF THE NUMBER OF SIGNALS
DETECTED FOR THE NONREDUNDANT AUDITORY,
NONREDUNDANT VISUAL, AND REDUNDANT SIGNAL
CONDITIONS

Source	<i>df</i>	<i>MS</i>	<i>F</i>
Periods (P)	5	6.971	6.77**
Signal Type (T)	2	38.270	31.68**
Subjects (S)	40	5.928	
P × T	10	1.346	2.71**
P × S	200	1.030	
T × S	80	1.208	
Residual (P × T × S)	400	.497	

** *p* < .01.

the auditory during the initial 30 minutes of watch. This effect is shown by the significant Signal × Period interaction. In spite of this reversal, individual comparisons between signal conditions demonstrated the significance of all three mean differences (*p* < .01).

Separate detection curves for the auditory (*R_a*) and visual (*R_v*) portions of the redundant signal are presented in Figure 2. Except for the initial drop in the auditory curve,

these two components of the redundant signal approximate the corresponding nonredundant auditory and visual functions. Total detection proportions for the two conditions bear out this similarity: *NR_a* = .78, *R_a* = .74, *NR_v* = .69, and *R_v* = .70.

The summation between the auditory and visual components of the redundant condition was estimated by computing the theoretical redundant functions for comparison with the observed. Assuming the independent detection probabilities for a visual signal (*P_v*) and an auditory signal (*P_a*), an estimate of the combined detectability (*P_{av}*) under a redundant signal condition is given in their joint probability, *P_aP_v* = *P_{av}*. It follows that the probability of a redundant detection is *P_a* + *P_v* - *P_{av}*, or $1 - (1 - P_a)(1 - P_v)$. Using the proportions of individual detections for the nonredundant presentations, a predicted redundant proportion within the 30-minute period was obtained for each *S* and then averaged for the group (*R'_{nr}*). Similarly, from the comparable detection proportions as separately recorded for the auditory and visual components of the bisensory signal condition,

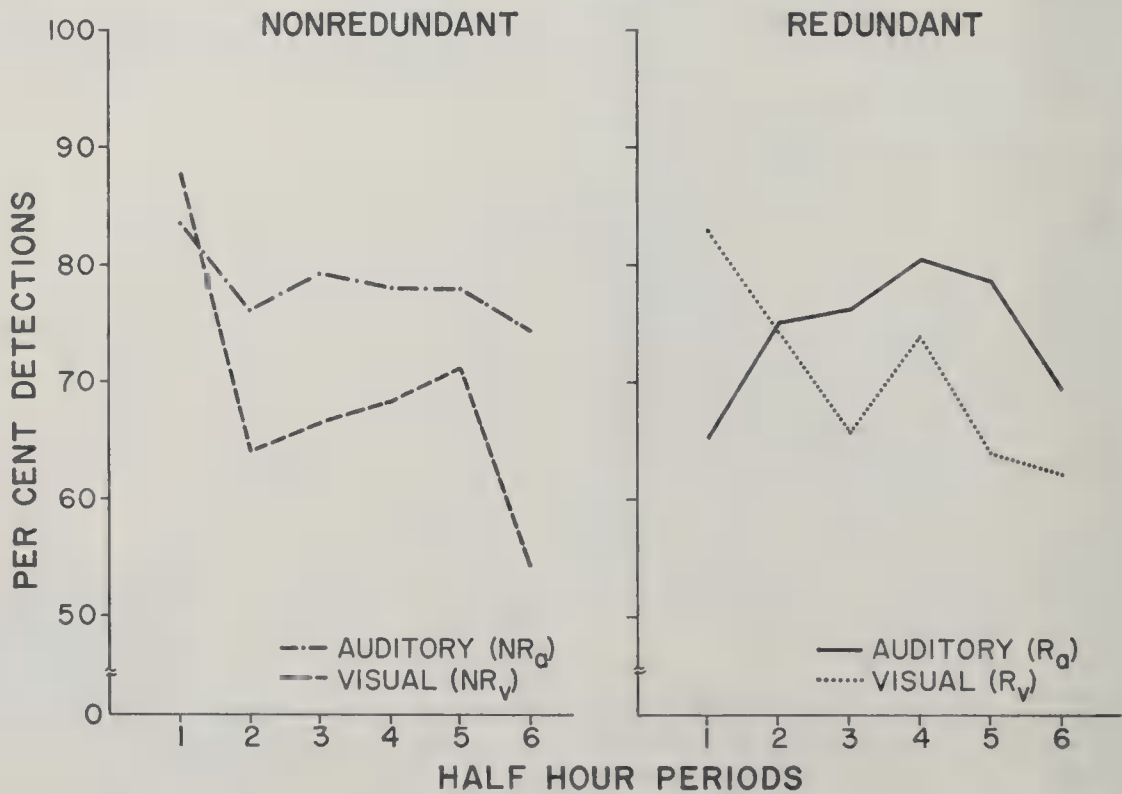


FIG. 2. Performance on the nonredundant auditory and visual compared with redundant auditory and visual signal types over the 3-hour vigilance session.

a second predicted function (R'_r) was computed. The resulting curves were plotted and then compared with the observed redundant function (Figure 3). Also shown in Figure 3 are plots of the percentage of combined or joint audio-visual detections for the redundant condition, and the corresponding joint probability curve.

While the similarity of both sets of curves is evident, the two predicted redundant curves constitute a slight though reliable overestimation of the observed curve over the last half of the vigilance session. An analysis of the number of predicted and observed detections (Table 2) showed this deviation was significant ($p < .05$). This effect is not shown, however, in the Curve \times Period interaction. In evaluating the main effect, note should be made of the restricted Curve \times Subject mean square—the error term for the F ratio. The magnitude of this interaction is inversely related to one measure of the accuracy of the procedure used in estimating the predicted redundant curves, viz., the between-subject consistency of directional differences in total number of detections for the three redundant conditions. As the extent of this consistency or reliability is considerable ($r = 1 - MS_{cs}/MS_s = .96$), the above mentioned F ratio

TABLE 2
ANALYSIS OF VARIANCE OF THE NUMBER OF PREDICTED OR OBSERVED SIGNAL DETECTIONS FOR THE RESPECTIVE REDUNDANT PERFORMANCE CURVES

Source	df	MS	F
Periods (P)	5	2.592	3.08*
Redundant Curves (C)	2	.506	3.44*
Subjects (S)	40	3.985	
P \times C	10	.131	1.28
P \times S	200	.842	
C \times S	80	.147	
Residual (P \times C \times S)	400	.102	

* $p < .05$.

provides a test that is extremely sensitive to any overall differences in elevation of the three redundant functions.

To evaluate the contributions of individual differences in ability to use the auditory and visual channels, the performance on each S 's preferred modality (as defined by the higher of his two nonredundant performance curves) was totaled for the group and the mean percentage of detection compared with the redundant function. These results also showed a mean detection rate for the bisensory condition significantly better than the "stronger" modality ($t = 4.44$, $df = 40$; $p < .01$).

DISCUSSION

These results both confirm and strengthen the summation hypothesis for redundant information. In view of these findings and previous results, there should now remain little doubt that an improvement in vigilance performance may be expected under a condition of bimodal signal presentation. Regardless of an individual monitor's preference for one particular sensory channel of a redundant signal, a significant informational contribution by the remaining, weaker signal component is evident. Thus, a recommendation for the adoption of dual channel displays whenever practicable in applied vigilance situations seems justified.

In theoretical terms, however, the data are less conclusive. The apparently reliable discrepancy over the last half of the watch session between the predicted and observed redundant functions limits the appropriateness

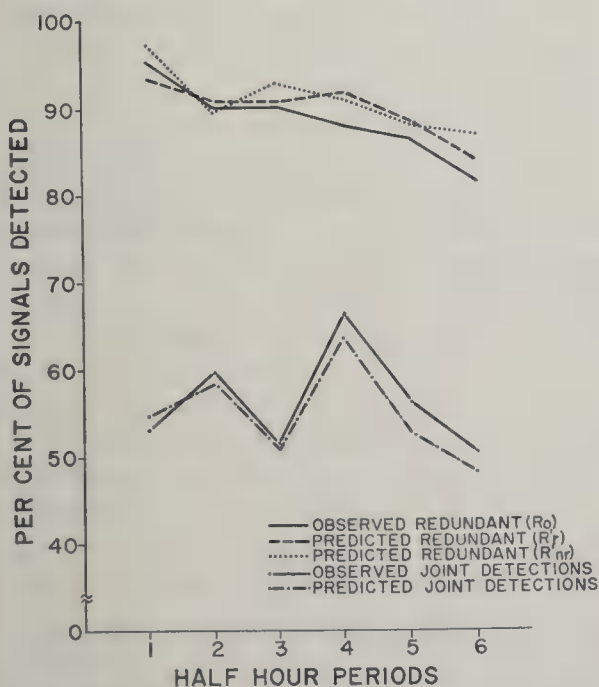


FIG. 3. Predicted performance compared with observed performance for the redundant signal condition.

of a summation hypothesis based entirely on the informational aspects of the component signals. While Buckner and McGrath (1961) assumed a similarly obtained overestimation to be due to a slight correlation in the number of total auditory and visual detections from which the predicted curve was calculated, this covariance factor was eliminated from the present analysis by obtaining predicted redundant scores by individual Ss. Another consideration is the possible influence of some intersensory interaction effect peculiar to the redundant conditions. Although the highly comparable redundant and nonredundant detection rates, when analyzed by separate modalities (Figure 2), may suggest the absence of such an effect, Curves R_a and R_v must be interpreted in conjunction with the number of joint audio-visual detections. That is, assuming the observed proportions of joint detections is high relative to the expected proportions, as estimated by $P_a P_v$ from the individual data in R_a and R_v , a redundant function predicted from the same data will tend to exaggerate the observed function. Reference to the combined detection curves in Figure 3 indicates this occurred. The Ss evidently had some difficulty in detecting both components of the bimodal signal at the be-

ginning of the watch. They did, however, show enough improvement to maintain a performance level greater than that to be expected from the detectabilities of the component signals during the last half of the session. Thus, it appears that the discrepancy between the predicted and observed redundant curves is due to this lack of decrement in the number of joint detections over the vigilance session.

REFERENCES

- BAKER, R. A., WARE, R. J., & SIPOWICZ, R. R. Signal detection in auditory, visual, and combined audio-visual vigilance tasks. *Canad. J. Psychol.*, in press.
- BUCKNER, D. N., & MCGRATH, J. J. A comparison of performances on a single and dual sensory mode vigilance task. Technical Report No. 8, 1961, Human Factors Research Incorporated, Los Angeles, California.
- LOVELESS, N. E. Signal detection with simultaneous visual and auditory presentation. Report No. 1027, 1957, Air Ministry, Flying Personnel Research Committee, London, England.
- MACKWORTH, N. H. Research on the measurement of human performances. *Med. Res. Council, Spec. Rep. Ser.*, 1950, No. 268.
- SCHAFER, T. H., & SHEWMAKER, G. A. A comparative study of the audio, visual, and audio-visual recognition differentials for pulses masked by random noise. *USN Electron. Lab. Rep.*, 1953, No. 372.

(Received May 22, 1962)

SELF-ESTEEM AND THE DIFFUSION OF LEADERSHIP STYLE¹

DAVID G. BOWERS

University of Michigan

Neither simple imitation, nor motivational coincidence, is adequate either to explain the frequent, superficial absence of similarities in leadership style across hierarchical levels (leadership climate) or to prescribe the best means for changing the style involved when climate does occur. Self-esteem of the lower-level supervisor is investigated as a mediating variable in this problem, in the context of an organization in which no formal human relations training had taken place. Variables were measured by questionnaires submitted to 17 foremen and their 330 male subordinates in a packaging materials plant. Hypotheses, all confirmed by the data, relate supportiveness of the foreman's supervisor to the foreman's behavior toward his subordinates through the attendant consequences of the foreman's self-esteem.

Observers frequently notice, and empirical investigators occasionally find, that supervisory styles tend to be similar between adjoining hierarchical levels within an organization. This resemblance Fleishman (1953) has termed "leadership climate." This finding quite naturally leads to speculation about how the similarity comes to exist, and how the style involved can effectively be altered.

One possible explanation is that supervisors at a lower level simply imitate in dealing with their subordinates the behavior and mannerisms of their own immediate superiors. This explanation assumes that such a course of action is perceived by the lower level supervisor as a path to success and reward.

An alternative explanation is that behavior stems from deep-seated personality and motivational forces, and that the similarities in supervisory behavior observed between hierarchical levels in organizations are simply due to the fact that supervisors at all levels of an organization tend to think alike because the high echelons selectively promote into the lower echelons those who think as they do.

These two explanations differ, not in their expectation of finding such a resemblance between hierarchical levels, but in their prescriptions for changing the style of the climate which exists. The imitation explanation

would emphasize a change in the behavior of the next higher supervisor, postulating that this change should then be observed to "trickle down" through the organization. The conative, or motivational, explanation would insist that a change at a high level would have little or no effect upon the behavior of supervisors at lower levels. Instead, some "therapeutic" program of conative change would be necessary at all levels.

This paper proposes that an integration of both these viewpoints provides, not only the best possible prescription for changing the style in those cases in which resemblances occur, but also a convenient means of explaining those situations in which, at first glance, resemblances do not seem to exist. In short, this paper proposes that the basis for diffusion of supervisory style is, in fact, hierarchical reward—both in the form of promotions and salary increases and in the form of expressions of approval and support from one's superior. However, a great deal of the contradictory evidence may be explained through conative elements which are at present too little stressed by leadership climate enthusiasts. Specifically, one such central process is proposed as the focus of this paper: foreman's self-esteem, defined as his general evaluation of his own worth.

RATIONALE

Whether a foreman's behavior toward his own subordinates patterns itself after that of

¹ This research was part of a larger project conducted at the Survey Research Center, University of Michigan, and was supported by the United States Air Force under Contract No. AF 49(638)-1032. It is listed as Technical Document No. AFOSR 2555.

his superior will depend, among other things, upon what he perceives that superior as wanting him to do. For that superior to behave considerably toward this foreman will not automatically result in the latter's behaving considerably in turn toward his own subordinates, unless the foreman perceives that this is what his superior wants him to do. In short, the superior's behavior toward the foreman is not necessarily an automatic one for imitation by the foreman in his actions toward his subordinates.

Let us assume, however, that the typical second-level supervisor feels rather indifferent toward whether the foreman behaves considerably toward his own subordinates, but, instead, is very much concerned about whether that foreman demonstrates the proper "initiative" by vigorously pressuring for production. That this attitudinal pattern is rather prevalent, at least within organizations whose work is repetitive, is suggested by Likert (1961, p. 81). Bass reaches a similar conclusion, that whereas a supervisor's subordinates will judge him in terms of his considerate behavior toward them, his superior will probably judge him in terms of the evidence of initiative which he demonstrates (Bass, 1960, p. 293). This is true, not only because that superior has different problems to cope with and different judgmental standards which result from those problems, but also because he has less opportunity to observe consideration than to observe initiative in a subordinate. Evidence of the latter is more readily obtained than evidence of the former.

For these reasons, we expect that the prevalent pattern of values regarding leadership practices will be fairly authoritarian or directive. Furthermore, the more this is the case, the less likely it is that a superior will support a foreman's needs for achievement or for affiliation in return for demonstrating considerate behavior toward his own subordinates, and the more likely that he will support the foreman in return for demonstrating a hard-headed program of "getting out the work."

This paper assumes, in addition, that non-economic, personality needs may be roughly categorized into two general areas, which we may label needs of "affiliation" and "achievement." It is further assumed that, at least

among those who have joined the ranks of management, motive structures are characterized by a greater importance for attainment of status, prestige, and recognition within the organization (achievement) than for friendly interaction with other persons (affiliation). Beyond this, it is assumed that, in the absence of any formal human relations training, the predominate value system of industrial management is one of a generally authoritarian or directive nature. This results in both the judgment of his effectiveness, and the communication of that judgment, to the foreman, not in terms of the consideration which he shows toward his subordinates, but in terms of his overt efforts to get out the work.

If, as we have assumed, foremen are more motivated personally to achieve than to affiliate (although by no means necessarily averse to the latter), it follows that their self-esteem—their judgments of themselves—will be in terms of the satisfactoriness which they perceive to exist regarding their performance. Since achievement is, in this situation, defined in terms of higher status, etc., it follows that those higher up in the organization, particularly the immediate superior, will be the chief source of information which results in these self-judgments.

When, for some reason, the superior communicates an unfavorable judgment to the foreman about the latter's work, the foreman is forced to re-evaluate himself in terms of this information. The communication tells him that he has not been rewarded—he may, in fact, conclude that he has been punished—and that the reason for this lack of reward is his behavior, which does not meet the expectations of his superior. Since his superior values initiative, and appears to care little about "consideration," the foreman is simultaneously presented with a lowering of his self-esteem and the groundwork for raising it to the former level.

What may this foreman do? He can conceivably reject the superior and his judgment, but this is unlikely, since it will involve some renunciation on his part of chances of fulfilling his aspirations. He can, alternatively, accept the criticism and take whatever action he feels is appropriate to bring his behavior into line with what his superior expects. He

may, instead, or in addition, raise his self-esteem by increasing his status or power in some other direction.

The last two solutions are not entirely incompatible and are, in fact, suggested to him by the situation. We may expect, and Bass (1960, p. 299) reviews evidence which supports this, that an individual will, in the absence of contrary forces, tend to generalize from the most salient source of opinion concerning him to other sources. In the present situation, therefore, we might expect that his superior's adverse judgment would lead him to believe that his subordinates feel similarly about him. Since individuals are quite likely to be more attracted to others who are perceived to approve of them, we may say that the foreman will alienate himself from his subordinates to the extent that he perceives this to be the case. As was outlined earlier, he will, to a considerable extent, be prevented from alienating himself from the superior because of the orientation of his motive structure.

Having alienated himself from his subordinates, the foreman's position is now more amenable to those actions which he must undertake to restore himself to favor with his superior. Having estranged himself from his subordinates, he is the more free to demonstrate initiative, pressure, criticize, and cajole these subordinates.

While he is doing this, and until the approval message raising his self-esteem occurs, the foreman remains with his self-esteem lowered. However, he can temporarily make the situation tolerable by recruiting other forms of respect: he can use the subordinate group for this purpose.

Brownfain (1952) leads us to believe that lower self-esteem persons are more dependent upon feedback from their social environment than are those higher in self-esteem. For this reason, the foreman feels that he can enhance his temporary image of himself by calling his subordinates together more often as a group than he would otherwise, with the intent of using this as an occasion within which to demonstrate his ability as an initiator and to receive recognition of this ability from his subordinates. As Andrews says, "The fact that he has the answers will enhance the

supervisor's standing in the eyes of his men" (Andrews, 1955; cited in Bass, 1960, p. 288).

However, as we observed earlier, he has alienated himself from his subordinates and is now, by emphasizing initiation in the group setting, behaving even less supportively toward them than before. It is very unlikely, therefore, that he will receive the desired recognition from his subordinates. Instead he will probably receive disapproval, resentment, and distrust—types of feedback unlikely to raise his self-esteem. Thus the original situation (lowered self-esteem, coupled with an attempt to alleviate this through use of group meetings) stands and is reinforced. The foreman, because of the limited possibilities open to him, repeats the cycle, probably until such time as it is interrupted by an expression of approval from his superior.

In short, we believe that if a foreman's superior expresses disapproval of the foreman's job efforts, this should lead to a decrease in the foreman's self-esteem. The latter should lead, in turn, under the conditions assumed, to an effort to retrieve himself temporarily and make the situation tolerable by calling his subordinates together more often, but less frequently for involving them in decision making. At the same time, his lowered self-esteem would lead to a generalized perspective of his subordinates' sharing the superior's unfavorable picture, which results in his alienating himself further from them, and, in addition to involving them less in decisions, to behaving less considerately toward them.

In other words, we conclude that the imitation of behavior alone is not sufficient to account for leadership climate. Instead, the values of both superior and subordinate, and the cognitive processes which occur in each regarding the other are elements essential to its explanation. The following illustration, using the designations "S" for second-level superior, "F" for foreman, and "W" for workers, will help clarify this conclusion:

Situation 1: S judges F in terms of F's pressuring W for production. He is considerate and full of praise for F when F does this, is not considerate and does not praise F when F does not. The former situation raises

F's self-esteem and thus rewards him for threatening behavior toward W. The latter situation threatens F for nonthreatening behavior. In this situation we would expect a negative correlation between the behavior of S and F.

Situation 2: S judges F in terms of F's consideration toward W, S is considerate toward F if F is considerate toward W in turn, and S is not considerate when F is not considerate toward W. Here we would expect a positive correlation between the behavior of S and F.

Situation 3: S judges F in terms of his coupling of consideration with enthusiasm for production. Since S will sometimes be attempting to get F's consideration back into line with his production emphasis (by behaving considerately toward him when he emphasizes production), and at other times attempting the reverse, the correlation between the behaviors of S and F should be very low or zero.

This paper assumes that Situation 1 is the most common, and that this is what the research site in the present study resembles, since the management does not appear to be deeply committed to an emphasis on human relations. Situation 2 probably occurs only rarely, and then in those instances in which only human relations training and principles have deliberately been emphasized. Situation 3 is that to be found in the "path-goal formulation," is the most effective, and is also probably rather rare (Georgopoulos, Mahoney, & Jones, 1957).

The preceding conclusions may be restated in hypothesis form as follows:

1. The less supportively the foreman's superior behaves toward him, the lower will be that foreman's self-esteem.

2. The lower the foreman's self-esteem, the more often he will take problems up with his subordinate group as a whole in the hope of raising his self-esteem, rather than for the purpose of solving the problems themselves.

3. The lower the foreman's self-esteem, the poorer he will perceive his subordinates' attitude toward him to be.

4. The attitude which he perceives his subordinates to hold will not necessarily be related at all to their actual attitude, but will

be positively related to the perceived supportiveness of his superior's behavior.

5. The poorer the attitude which the foreman perceives his subordinates to share with his superior concerning him, the more he will alienate himself from his subordinates.

6. The more alienated from his subordinates the foreman is, the less he will behave supportively toward them, and the less he will accept their opinions and advice in solving problems and making decisions.

METHOD

The data used in this study were obtained by questionnaires administered in December 1958 to 17 foremen and their 330 male subordinates in two plants of a company manufacturing packaging materials.² The variables measured are described below, together with an example of each. For the purposes of this report, scale directions were adjusted to reflect a high degree of the variable as named. Ratings of foremen were obtained from their subordinates; ratings of each foreman's supervisor were obtained from the foremen themselves.

Supportiveness

A five-item measure of the extent to which a supervisor or foreman is understanding of the needs of his subordinates, encourages them toward self-improvement in terms of promotions, provides praise for jobs well done, and shows warm and friendly concern for their welfare.

Example. My foreman spends hardly any time helping his people work themselves up to a better job by showing them how to improve their performance.

- (1) Strongly agree
- (2) Agree
- (3) Undecided
- (4) Disagree
- (5) Strongly disagree

Group Approach

A three-item measure of the extent to which decisions are made, or problems are handled, by the foreman by calling together the subordinates as a group.

Example. If your foreman finds a new method for doing a certain part of the work in the department, what does he do?

- (1) He usually takes it up *separately* with the persons directly involved.
- (5) He usually takes it up with the employees *as a group*.

² Women were omitted from the groups used in this analysis, since their situation contained some peculiarities which it was felt would cloud the issue under discussion.

Acceptance of Advice

A four-item measure of the extent to which the foreman accepts the opinions and advice of his subordinates in making decisions or solving problems.

Example. If complaints are received from another department or the customer about the inferior quality of the work done in your department, what does the foreman do?

- (1) He almost never accepts the opinion of the employees.
- (2) He sometimes accepts it.
- (4) He usually accepts it.
- (5) He almost always accepts the opinion of the employees.

Self-Esteem

The foreman's opinion of himself, obtained through his mean score of 10 pairs of adjectives arranged into a semantic differential.

Example. In the following question, we would like you to think about yourself and answer in terms of what words best describe you. For example, look at the first line of boxes. A man who feels important should put his check mark (✓) close to that word. If, on the other hand, he feels he is rather unimportant, he should put his check mark close to that word. If he feels, however, that he falls somewhere in the middle, he should put his check mark somewhere on the line where he feels it best fits him.

(1)	(2)	(3)	(4)	(5)	
Important					Unimportant

Perceived Attitude toward Foreman

A single-item measure of overall satisfaction with the foreman, rated both by his subordinates and by him in terms of what the respondent felt was the average attitude of his subordinates toward him.

Example. We would like you to answer the following questions as you think an average employee in your department would answer them.

All in all, how satisfied are you with your foreman?

- (1) Very satisfied with my foreman
- (2) Quite satisfied
- (3) Fairly satisfied
- (4) A little dissatisfied
- (5) Very dissatisfied with my foreman

Alienation from Subordinates

A score representing the difference between the foreman's attitude toward other persons on his level on four items and his attitude toward his subordinates on six comparable items.

Example. To what extent do you feel that you and the men you supervise belong to a team that works together?

versus

To what extent do you feel that you and the men on your level belong to a team that works together?

- (1) To a very great extent I feel we belong to a team
- (2) To a considerable extent
- (3) To some extent
- (4) To a very little extent
- (5) To no extent

Need for Achievement

The foreman's mean response to five items which ask for the importance which he attaches to opportunities for personal accomplishment.

Example. How important is it for you to do your best in whatever you undertake?

- (1) Very important
- (2) Quite important
- (3) Somewhat important
- (4) Of little importance
- (5) Of no importance

Need for Affiliation

The foreman's mean response to six items which ask for the importance which he attaches to opportunities for warm interpersonal relations.

Example. How important is it to you to feel that others like you?

- (1) Very important
- (2) Quite important
- (3) Somewhat important
- (4) Of little importance
- (5) Of no importance

All five hypotheses were tested by computing Spearman rank-difference correlation (ρ) coefficients between the measures concerned. For 17 cases, a coefficient of .41 is significant at the .05 level of confidence (one-tailed test); a coefficient of .58 is significant at the .01 level of confidence.

RESULTS

Before presenting the results which test the hypotheses, evidence must be presented to support the contention that the situation fits the problem as stated. First of all, we assume that, in general, foremen are more achievement than affiliation oriented. The data indicate that, of the 17 foremen, 13 present this pattern. A sign test indicates that such a split in proportions is adequate to demonstrate this point. Secondly, we assume that supervisors and foremen had not been given human relations training; experience and interviews demonstrated to the satisfaction of the investigator that there had been no such training

in this organization prior to the time this study was done, and that foremen and supervisors had little formal knowledge of human relations principles. Third, it is desirable that our situation fit those in which leadership climate evidence has been lacking: we prefer that the supportiveness which the foreman receives from his supervisor not correlate with the supportiveness that he accords his subordinates. That is, in fact, the case. The two measures correlate only .22, which is not significant.

The first hypothesis leads us to expect a positive correlation between the supervisor's supportiveness and the foreman's self-esteem; a positive coefficient of .80 confirms this hypothesis.

The second hypothesis tells us that a negative correlation should exist between the foreman's self-esteem and the extent to which he employs a group approach. This hypothesis is also confirmed: a negative correlation ($-.50$) results.

The third hypothesis states that a positive relationship should result between the foreman's self-esteem and his perception of his subordinates' attitude toward him. A positive coefficient of .66 confirms this.

A fourth hypothesis states that the foreman's perception of this attitude should bear no necessary relation to the attitude itself as expressed by those subordinates, but should correlate positively with the supportiveness of the foreman's supervisor. These aspects of the hypothesis are also confirmed: the foreman's perception of the attitude of his subordinates toward him correlates only $-.09$ with the actual attitude expressed by those subordinates, whereas it correlates positively (.73) with his supervisor's supportiveness.

According to the fifth hypothesis, the poorer the attitude which the foreman perceives to exist toward him among his subordinates, the more he attitudinally alienates himself from those subordinates. As expected, we find a negative correlation ($-.50$) between the two measures.

The last hypothesis states that alienation

from his subordinates should be related to his behaving less supportively toward them and accepting less their opinions and advice in solving problems and making decisions. This hypothesis is also confirmed: alienation correlates $-.67$ with the foreman's rated supportiveness and $-.64$ with his rated acceptance of opinions and advice.

CONCLUSION

The results just presented provide strong evidence to support the position outlined at the outset. Evidence to support the simplest statement of the leadership climate concept frequently does not appear because certain central processes intervene between behavior at one level and behavior at the next level down in the hierarchy. Particularly, one such conative mediator is the foreman's self-esteem which aids in translating for him the behavior of his superior into mandates for his own actions. In short, leadership climate is not, as many treatments might prefer, a simple imitation at a lower level of a pattern practiced at the next level up in the hierarchy, nor is it simply a reflection of the subject's values and motives. Instead, it is a matter of perceived selective reward, mediated by the cognitive and conative structures of the lower-level individual.

REFERENCES

- ANDREWS, R. E. *Leadership and supervision: A survey of research findings*. (Personnel Management Series No. 9) Washington, D. C.: United States Government Printing Office, 1955.
- BASS, B. M. *Leadership, psychology, and organizational behavior*. New York: Harper, 1960.
- BROWNFAIN, J. J. Stability of the self-concept as a dimension of personality. *J. abnorm. soc. Psychol.*, 1952, 47, 597-606.
- FLEISHMAN, E. A. The leadership opinion questionnaire. In R. M. Stogdill and A. E. Coons (Eds.), *Leadership behavior: Its description and measurement*. Columbus: Ohio State University, Bureau of Business Research, 1957.
- GEORGOPOULOS, B., MAHONEY, G., & JONES, N. A path-goal approach to productivity. *J. appl. Psychol.*, 1957, 41, 345-353.
- LIKERT, R. *New patterns of management*. New York: McGraw-Hill, 1961.

(Received May 22, 1962)

JOB ATTITUDES IN MANAGEMENT:

II. PERCEIVED IMPORTANCE OF NEEDS AS A FUNCTION OF JOB LEVEL ¹

LYMAN W. PORTER

University of California, Berkeley

By means of a questionnaire, 1916 managers indicated the degree of importance they attached to 13 items representing 5 areas of psychological needs. Respondents represented all levels of management and many different types of companies. The 5 need areas studied were Security, Social, Esteem, Autonomy, and Self-Actualization. Results showed that there was some relationship between vertical level of position within management and degree of perceived importance of needs. Higher-level managers placed relatively more emphasis on Self-Actualization and Autonomy needs than did lower-level managers. For each of the other 3 types of needs, however, there were no differences between responses from higher-level vs. lower-level managers. The findings from this study were compared with those from recent related studies.

In a recent study (Porter, 1962) perceived deficiencies in need fulfillment were examined as a function of five levels of management. Results of that research showed that the vertical level of position within management had an important effect on the degree of perceived need satisfaction: managers in lower level positions reported less satisfaction than those at higher levels. The present study investigates the same set of need satisfaction questions, but is focused on the perceived importance of these needs to the individual rather than on the perceived satisfaction or fulfillment. Again, as in the previous study, the results are analyzed as a function of management level.

Two other previous studies have some bearing on the present investigation. One of these studies (Porter, 1961), utilizing a Maslow-type conceptualization of types of needs, reported results on perceived importance of needs for the two lowest levels of management within three different organ-

izations. In that study, which used the same questionnaire items employed in the present study, it was found that on most of the specific need questions the two levels of management did not differ significantly in the amount of importance they attached to the needs. However, there was a trend for the lowest level managers, the first level supervisors, to attach slightly more importance to most of the items compared with the managers at the level immediately above them. Somewhat the same results were also found in a recent study by Rosen and Weaver (1960). Although their results were based on an entirely different set of questions concerning conditions of work, they also found relatively small differences among the three lowest levels of management within a single plant. In those few instances where significant differences were found, first-level supervisors regarded specific conditions as more important than higher-level managers. Another finding common to both the Porter and the Rosen and Weaver studies was that in each case the same pattern of importance among the items investigated existed across the different management levels. That is, in each study, the specific items that one management level considered most important were usually the same items that the other management levels also considered most important. Rosen and Weaver (1960) have interpreted this latter finding as showing

¹ This study was carried out as part of the research program of the Institute of Industrial Relations, University of California, Berkeley. It was started while the author was a Ford Foundation Faculty Research Fellow. The Institute of Social Sciences at the University of California and the American Management Association contributed to the support of the research assistance.

The author is indebted to Mildred Henry, Larry Stewart, and Robert Andrews for assistance in tabulation of the data.

that at least for the lowest levels of management "perhaps one can accurately talk about 'management' as a meaningful, cohesive class sharing common motivations re what they want from their work . . ." (p. 392).

The present study, with a large sample of respondents representing all levels of management in many different types of organizations, seeks first to extend the results of these earlier Porter and Rosen and Weaver studies of "importance" carried out on samples of limited size in lower levels of management. Secondly, this study focuses especially on a comparison of the results for importance to the results for need fulfillment deficiencies previously obtained from the same sample of respondents (Porter, 1962). In essence, this latter comparison investigates the question of whether there are systematic changes in the importance attached to various needs as one goes from lower to higher levels of management, as is the case when need fulfillment deficiencies are examined.

METHOD

Questionnaire

The data for this study were collected by means of a questionnaire described in detail in previous articles (Porter, 1961, 1962). The data are based on answers to parts of 13 items contained in the questionnaire. All 13 items are relevant to a Maslow-type need hierarchy system. A sample item, as it appeared in the questionnaire, was as follows:

The opportunity for independent thought and action in my management position:

- (a) How much is there now?
(min) 1 2 3 4 5 6 7 (max)
- (b) How much should there be?
(min) 1 2 3 4 5 6 7 (max)
- (c) How important is this to me?
(min) 1 2 3 4 5 6 7 (max)

Only the answers to Part c on the importance of each need are analyzed in this study. A previous paper (Porter, 1962) dealt with the answers to Parts a and b of each item.

Categories of Needs and Specific Items

Listed below are the hierarchical categories of needs studied in this investigation, along with the specific items used to elicit information on each category. The items were randomly presented in the questionnaire, but are here listed systematically according to their respective need categories. The categorization system has been described in detail in a previous paper (Porter, 1961). Essentially, it is based on Maslow's system of classifying different needs ac-

cording to their prepotency of elicitation (Maslow, 1954). The categories and their specific items follow:

I. Security needs

1. The *feeling of security* in my management position

II. Social needs

1. The *opportunity, in my management position, to give help to other people*
2. The *opportunity to develop close friendships* in my management position

III. Esteem needs

1. The *feeling of self-esteem* a person gets from being in my management position
2. The *prestige* of my management position *inside* the company (that is, the regard received from others *in* the company)
3. The *prestige* of my management position *outside* the company (that is, the regard received from others *not* in the company)

IV. Autonomy needs

1. The *authority* connected with my management position
2. The *opportunity for independent thought and action* in my management position
3. The *opportunity, in my management position, for participation in the setting of goals*
4. The *opportunity, in my management position, for participation in the determination of methods and procedures*

V. Self-Actualization needs

1. The *opportunity for personal growth and development* in my management position
2. The *feeling of self-fulfillment* a person gets from being in my management position (that is, the feeling of being able to use one's own unique capabilities, realizing one's potentialities)
3. The *feeling of worthwhile accomplishment* in my management position

Procedure and Sample

As was described previously (Porter, 1962), the sample of 1,916 respondents was obtained by mailing the questionnaire to nearly 6,000 managers and executives.² About two-thirds of the sample came from manufacturing companies, 7% from transportation and public utilities, 7% from finance and insurance, 5% from wholesale and retail trade, and the remaining 15% from among other types of companies. It should also be noted that the wide distribution of the questionnaire provided a sample that consisted of one or a few individuals from each of many different companies located throughout the country.

A number of personal data items were included at the end of the questionnaire to permit the

² The assistance of the American Management Association, and particularly Robert F. Steadman, in obtaining the sample of respondents is gratefully acknowledged.

TABLE 1

DISTRIBUTION OF *N* OF TOTAL SAMPLE BY FIVE MANAGEMENT LEVELS AND FOUR AGE GROUPS
AND CHARACTERISTICS OF SAMPLE BY MANAGEMENT LEVEL

Management level	Age group				Total <i>N</i> for level	Median age	College degree (%)
	20-34	35-44	45-54	55+			
President	7	31	51	25	114	48.2	76.1
Vice President	52	240	212	107	611	45.5	72.6
Upper-Middle	95	288	206	70	659	43.1	75.0
Lower-Middle	100	208	98	25	431	40.6	75.0
Lower	32	46	14	9	101	39.2	76.2

classification of managers on a number of independent variables. The two relevant ones for this study were level of position with management, and (as a control variable) age of the respondent. A five-category system was used to classify the respondents according to level of position. The rationale and procedure of this classification scheme were explained in the previous paper (Porter, 1962). The five management-level categories were: President, Vice President, Upper-Middle, Lower-Middle, and Lower. Although level of position is the major independent variable in this study, as in the previous one, it was felt necessary to classify respondents by age as well as by management level. In this way, the age variable, which is to some extent correlated with level of position, could be held constant. Also, this permitted replication of results within several independent groups. The four age categories were: 20-34, 35-44, 45-54, and 55+.

Table 1 presents the number of respondents in each management level within each age group. This table, which can be referred to in determining the *N*s for the subgroups of the subjects in Tables 2, 3, and 4, shows that higher-level managers were somewhat older than lower-level managers. The table also shows that with regard to the amount of formal education of the managers, all levels were approximately equivalent. This fact will be relevant later to the interpretation of the results of the study. Since the lowest level of management in this study had a higher percentage of college graduates than might be expected (due to the fact that it was probably composed mostly of office-type supervisory or staff personnel), any conclusions involving this level would not necessarily apply to first-line foremen who have not had as much formal schooling.

RESULTS

The basic results of this study are presented in Table 2. In this table, the mean response to Part c ("How important is this to me?") of each item is presented for each subgroup of respondents, classified by management level and by age. Since the scale for Part c of each item runs from 1 for

minimum importance to 7 for maximum importance, the larger an entry in a given cell of Table 2 the greater the importance attached to that item by a subgroup of respondents. This table shows that for the Autonomy and Self-Actualization areas there was a general trend for greater importance to be attached to these needs at the higher levels of management. Also, Table 2 shows that the items in the Autonomy and Self-Actualization areas elicited the greatest expression of importance. These latter trends will become more apparent in Table 4.

Table 3 is designed to summarize more clearly the changes in size of mean importance for each item from higher to lower levels of management. Whenever the mean values for two successive levels of management differed by more than .05 scale units, the change was counted as an increase (+) if the mean for the next lower level was larger, and was counted as a decrease (−) if smaller. Where separations of means were equal to, or less than, .05 scale units, they were counted as no changes (O's). (The decision to use .05 scale units as the dividing line between "changes" and "no changes" was an arbitrary one intended to allow for trends to appear in the data while at the same time reducing the effects of chance fluctuations among the means.) The numbers of each type of change—increases, decreases, and equalities—in size of mean importance that occur between the top and bottom levels of management are shown in the columns of Table 3. As an example, for Esteem Item III-1 ("The feeling of self-esteem a person gets from being in my management position"), Table 3 shows that in the 20-34 age group mean importance in-

creased once and decreased twice, going from higher to lower levels of management; in the 35-44 age group there were two increases, one no change and one decrease between top and bottom level managers; in the

45-54 group there were two increases and one decrease; and in the 55+ group there were one increase and two no changes. Summing across age groups for Item III-1 shows that in six instances there were increases

TABLE 2
MEAN IMPORTANCE OF EACH NEED CATEGORY ITEM WITHIN EACH SUBGROUP OF RESPONDENTS

Need category	Item	Age group	Management levels				
			President	Vice President	Upper-Middle	Lower-Middle	Lower
Security	I-1	20-34	—	4.94	4.62	5.06	4.84
		35-44	5.59	5.31	5.13	5.17	5.39
		45-54	5.73	5.58	5.35	5.63	—
		55+	5.60	5.71	5.86	5.92	—
Social	II-1	20-34	—	5.92	5.86	5.71	5.37
		35-44	6.34	6.12	6.03	6.00	6.17
		45-54	6.16	6.33	6.23	6.06	—
		55+	6.04	6.49	6.43	6.36	—
	II-2	20-34	—	3.87	4.27	4.71	4.22
		35-44	4.53	4.53	4.47	4.57	4.78
		45-54	4.53	4.84	4.43	4.77	—
		55+	4.96	5.09	5.19	5.52	—
	III-1	20-34	—	5.13	5.34	5.26	5.09
		35-44	5.12	5.25	5.33	5.24	5.20
		45-54	4.94	5.16	5.09	5.20	—
		55+	5.20	5.45	5.41	5.40	—
	III-2	20-34	—	5.21	5.32	5.43	4.91
		35-44	5.34	5.44	5.43	5.42	5.48
		45-54	5.51	5.50	5.29	5.44	—
		55+	5.72	5.53	5.53	5.76	—
	III-3	20-34	—	5.04	5.16	5.04	4.44
		35-44	5.34	5.35	5.29	5.03	5.24
		45-54	5.22	5.30	4.92	5.11	—
		55+	5.20	5.38	5.44	5.52	—
Autonomy	IV-1	20-34	—	6.00	5.58	5.51	5.12
		35-44	6.06	5.83	5.81	5.60	5.52
		45-54	6.02	5.85	5.39	5.35	—
		55+	6.24	5.81	5.61	5.28	—
	IV-2	20-34	—	6.37	6.43	6.18	6.03
		35-44	6.66	6.42	6.31	6.14	6.15
		45-54	6.35	6.42	6.09	5.94	—
		55+	6.36	6.33	6.10	5.88	—
	IV-3	20-34	—	6.04	6.13	5.73	5.34
		35-44	6.72	6.30	6.10	5.91	5.67
		45-54	6.35	6.42	5.92	5.80	—
		55+	6.32	6.30	5.91	5.96	—
	IV-4	20-34	—	5.63	5.63	5.30	4.97
		35-44	5.47	5.79	5.71	5.63	5.54
		45-54	5.04	5.92	5.55	5.55	—
		55+	5.80	5.79	5.87	5.68	—

Table 2—Continued

Need category	Item	Age group	Management levels				
			President	Vice President	Upper-Middle	Lower-Middle	Lower
Self-Actualization	V-1	20-34	—	6.63	6.67	6.57	6.47
		35-44	6.72	6.57	6.46	6.33	6.50
		45-54	6.33	6.12	6.04	6.04	—
		55+	6.24	5.85	5.87	5.88	—
	V-2	20-34	—	6.21	6.34	6.13	6.28
		35-44	6.66	6.37	6.33	6.24	6.39
		45-54	6.43	6.41	6.21	6.17	—
		55+	6.32	6.50	6.29	6.28	—
	V-3	20-34	—	6.33	6.47	6.22	6.03
		35-44	6.81	6.51	6.44	6.28	6.48
		45-54	6.55	6.54	6.38	6.28	—
		55+	6.20	6.57	6.36	6.32	—

Note.—Measures of variability are not included in this table. However, the standard deviations for each mean were computed and they showed the following distribution by quartiles: $Q_1 = .90$; $Q_2 = 1.22$; $Q_3 = 1.40$.

TABLE 3
NUMBER OF CHANGES IN SIZE OF MEAN IMPORTANCE FROM HIGHER TO LOWER
LEVELS OF MANAGEMENT WITHIN FOUR AGE GROUPS

Need category	Item	Age group												Total sample		
		20-34			35-44			45-54			55+					
		+	0	-	+	0	-	+	0	-	+	0	-	+	0	-
Security	I-1	1	0	2	1	1	1	1	0	2	3	0	0	3	0	0
Social	II-1	0	0	3	1	1	2	1	0	2	1	0	2	3	1	9
	II-2	2	0	1	2	1	1	2	0	1	3	0	0	9	1	3
Category total		2	0	4	3	2	3	3	0	3	4	0	2	12	2	12
Esteem	III-1	1	0	2	2	1	1	2	0	1	1	2	0	6	3	4
	III-2	2	0	1	2	2	0	1	1	1	1	1	1	6	4	3
	III-3	1	0	2	1	1	2	2	0	1	3	0	0	7	1	5
Category total		4	0	5	5	4	3	5	1	3	5	3	1	19	8	12
Autonomy	IV-1	0	0	3	0	1	3	0	1	2	0	0	3	0	2	11**
	IV-2	1	0	2	0	1	3	1	0	2	0	1	2	2	2	9 ^a
	IV-3	1	0	2	0	0	4	1	0	2	0	2	1	2	2	9 ^a
	IV-4	0	1	2	1	0	3	1	1	1	1	1	1	3	3	7
Category total		2	1	9	1	2	13	3	2	7	1	4	7	7	9	36**
Self-Actualization	V-1	0	1	2	1	0	3	0	1	2	0	2	1	1	4	8*
	V-2	2	0	1	1	1	2	0	2	1	1	1	1	4	4	5
	V-3	1	0	2	1	0	3	0	1	2	1	1	1	3	2	8
Category total		3	1	5	3	1	8	0	4	5	2	4	3	8	10	21*
Total all items		12	2	25	13	10	29	12	7	20	15	11	13	52	30	87**

^a Approaches significance ($p = .10$).
* $p = .05$.
** $p = .01$.

TABLE 4
RANKS OF MEAN IMPORTANCE FOR FIVE NEED CATEGORIES
WITHIN EACH SUBGROUP OF RESPONDENTS

Management level	Age group	Need category					Self-actualization
		Security	Social	Esteem	Autonomy		
President	20-34	—	—	—	—		—
	35-44	3	4	5	2		1
	45-54	3	4	5	2		1
	55+	3	4	5	2		1
Vice President	20-34	4	5	3	2		1
	35-44	5	4	3	2		1
	45-54	3.5	3.5	5	2		1
	55+	4	3	5	2		1
Upper-Middle	20-34	5	4	3	2		1
	35-44	5	4	3	2		1
	45-54	3	4	5	2		1
	55+	3	4	5	2		1
Lower-Middle	20-34	5	4	3	2		1
	35-44	5	3	4	2		1
	45-54	3	4	5	2		1
	55+	3	2	5	4		1
Lower	20-34	3	5	4	2		1
	35-44	4	3	5	2		1
	45-54	—	—	—	—		—
	55+	—	—	—	—		—

in perceived importance between one level of management and the next lower level, three instances of no change, and four cases of decreases. Thus, for this particular item, increases and decreases in importance were nearly balanced between higher and lower levels of management.

Table 3 shows, first, that there was an overall trend (significant at the .01 level) for higher-level managers to attach greater importance than lower-level managers to the needs studied in this investigation. However, this overall trend was provided almost entirely by one item in the Social need category, three in the Autonomy category, and two in the Self-Actualization category. Only 4 of the 13 items exceeded or approached significance by the sign test. It is apparent then, from Table 3, that managers at the higher echelons regarded need fulfillment in their jobs as more important only for certain questions, mostly in the Autonomy and Self-Actualization areas.

Several other findings in Table 3 can be noted. The 55+ age group deviated some-

what from the other three age groups in terms of some of the specific trends. For example, this group's trend was reversed on the Security item, and for the three Self-Actualization items its trend was very weak compared to that evident in the other three age groups. Another finding seen in Table 3 is the difference in trends for the two items in the Social need category. Item II-1 ("The opportunity to give help to other people") becomes increasingly more important at the higher levels of management, while item II-2 ("The opportunity to develop close friendships") becomes increasingly less important at higher levels of management.

Table 4 presents the ranking of mean importance for the five need categories within each subgroup of respondents. The category values on which these rankings are based were computed on the mean importance of the separate need items within each category. From this table it is possible to see how the five categories of needs were ranked in importance by each age group within each management level. For example, for the Vice Pres-

ident level within the 35–44 age group, Self-Actualization needs ranked first in importance, followed by Autonomy, Esteem, Social, and Security needs.

Table 4 shows that Self-Actualization and Autonomy needs were ranked first or second by every subgroup of subjects except one. This table also shows that Security and Social needs are about tied for third place and that they are fairly close to Esteem needs which ranked fifth for all of the subgroups combined. Table 4 also shows that no essential differences existed among respondents at different management levels in terms of how they ranked the relative importance of the five categories of needs. However, there were some differences in rankings as a function of age. For example, Security needs ranked relatively higher among the five categories of needs for the older respondents, while Esteem needs tended to increase their rank among the younger managers.

DISCUSSION

The results presented in the preceding section can be discussed in relation to several recent relevant studies. First, they can be compared with those obtained on need importance in a previous study using the same set of questions (Porter, 1961). The earlier investigation showed that the lowest level of management considered most of the need items as slightly more important than did the managers immediately above them. In contrast, in the present study there was no tendency in this direction and even a reverse trend for some items. That is, the present study showed that the higher the level of management, the greater was the importance attached to some of the need items, especially those in the Autonomy and Self-Actualization categories. It should be emphasized that comparisons between the results for the two studies can be made at only the two lowest levels of management—Lower and Lower-Middle—since these were the only levels investigated in the earlier study. For these lower levels, the previous study used a sample drawn from only three companies, whereas the present study drew a nationwide sample from a wide variety of companies. The other major difference in the samples is the fact that in the earlier study

the respondents in the higher of the two levels possessed somewhat more formal education than did those in the lower level, while in the present study there was no difference in amount of education between the two levels. It is possible that when questions are asked about the importance of various needs and opportunities in the work situation, those with less formal education may have a stronger response set to consider “everything” as quite important. On the other hand, those with more education may be more selective and restrained in using attitude scales and in responding to such questions. Therefore, the control of the education factor in the present study makes it possible to assess the independent effects of level of position more accurately than was the case in the first investigation.

In other respects, the former study and the present study provided similar findings, especially in regard to the ranking of the five need categories. In both studies, among the five need areas, Self-Actualization needs ranked highest in relative importance and Social and Esteem needs ranked lowest.

The findings obtained in the present study also permit comparison with the Rosen and Weaver study (1960) of perceived importance of working conditions within management. Their study, using respondents from a single plant, found as did the present study, that managers at different organizational levels tended to produce similar rank orders of the items studied. However, they also found a tendency for the lowest level of management to consider a number of items as more important than did higher-level managers. This again is contrary to the present findings obtained from a larger, more diverse sample. The differences in results may be due to the particular conditions existing in the single plant studied by Rosen and Weaver and also could be accounted for by possible differences in amount of formal education among their managerial levels. In addition, of course, the Rosen and Weaver study was concerned with one area of job attitudes, conditions of work, while the present study investigated a different area, psychological needs.

Finally, the most interesting comparison of the results of the present study on need importance is with the results on need ful-

fillment deficiencies obtained from the identical sample of managers responding to the identical need items (Porter, 1962). In general, the current results for importance show less definite trends of systematic changes from higher to lower levels of management, when compared with the results for need fulfillment deficiencies. Fewer of the specific need items reached or approached a statistical significance level in registering consistent changes from one level of management to the next for the importance data. That is, there was more homogeneity among different levels of managers in terms of the degree of importance they attached to psychological needs than there was in terms of the degree of perceived satisfaction they ascribed to these same needs. This is exactly what would be expected if answers concerning the importance of different needs can be assumed to reflect individual preferences of the respondent as well as the duties and responsibilities of his position. Answers to questions concerning perception of satisfactions provided by a job position should be less influenced by individual differences, and therefore more systematic differences among management levels would be expected in the area of perceived satisfaction than in the area of importance.

A more specific comparison of the sets of data for importance and satisfaction shows that while perceived need fulfillment deficiencies were clearly larger at the lower levels of management, perceived importance of at least two areas of needs was somewhat less at these levels compared with higher levels. Thus, as one goes down the management hierarchy, Autonomy and Self-Actualization needs become much less fulfilled, but these same two types of needs are also seen as somewhat less important in the lower management echelons. A clear exception, however, involves the Esteem needs, which were perceived as having greater deficiencies in fulfillment at the lower management levels but which were considered as important at these levels as at higher levels. Both the Security and Social categories tended to show no consistent changes in either deficiencies or importance from one management level to another.

Both the previous need deficiency study

and the present importance study have one striking result in common: the saliency of Self-Actualization and Autonomy needs. Comparisons among the five categories of needs showed that Self-Actualization and Autonomy were the needs perceived as least often fulfilled and also the needs felt to be the most important. These findings held up consistently among almost all subgroups of respondents at all ages and management levels.

CONCLUSIONS

The results of the present study show that the level of a position or job within management is related to the degree of perceived importance of needs: higher-level managers tend to regard certain types of needs—chiefly, Autonomy and Self-Actualization needs—as more important to them in their jobs than do lower-level managers in their jobs. However, this overall relationship between job level and the differential importance of needs appears to be definitely less strong than the previously reported relationship between job level and the degree of perceived fulfillment of needs. The present results also indicate, though, that all five levels of management tend to be similar in the relative ranks they give to the importance of the five different need areas. A final implication, based on both the results of the present study and on previously reported results, is that since Self-Actualization and Autonomy needs are seen by all management levels as the most important and least fulfilled types of needs, they are probably the most critical psychological need areas for organizations to consider in their relations with their managers and executives.

REFERENCES

- MASLOW, A. H. *Motivation and personality*. New York: Harper, 1954.
- PORTER, L. W. A study of perceived need satisfactions in bottom and middle management jobs. *J. appl. Psychol.*, 1961, **45**, 1-10.
- PORTER, L. W. Job attitudes in management: I. Perceived deficiencies in need fulfillment as a function of job level. *J. appl. Psychol.*, 1962, **46**, 375-384.
- ROSEN, H., & WEAVER, C. G. Motivation in management: A study of four managerial levels. *J. appl. Psychol.*, 1960, **44**, 386-392.

(Received May 22, 1962)

RETRANSLATION OF EXPECTATIONS: AN APPROACH TO THE CONSTRUCTION OF UNAMBIGUOUS ANCHORS FOR RATING SCALES¹

PATRICIA CAIN SMITH AND L. M. KENDALL

Cornell University

A procedure was tested for the construction of evaluative rating scales anchored by examples of expected behavior. Expectations, based on having observed similar behavior, were used to permit rating in a variety of situations without sacrifice of specificity. Examples, submitted by head nurses as illustrations of nurses' behavior related to a given dimension were retained only if reallocated to that dimension by other head nurses, and were then scaled as to desirability. Agreement for a number of examples was high, and scale reliabilities ranged above .97. Similar content validity should be obtained in other rating situations.

In many situations, the use of ratings as criteria for validation of tests and as indices of effectiveness of educational, motivational, and situational changes, involves extreme demands upon the quality of the ratings. Ratings from different raters in different situations should be really equivalent since they are almost always treated as if they were so. This demand for comparability means that interpretation of the rating must not deviate too widely from rater to rater or occasion to occasion, either in level (evaluation) or in dimension (trait, situational characteristic, job demand, temporal requirement, etc.). The present report covers the development of a rationale for a series of scales with such characteristics, and the testing of a procedure for their construction.

Psychologists seeking to establish reliable and valid rating systems have tended to impose their own values, interpretations, and beliefs about behavior upon the raters. Those who believe in trait theory construct scales based on presumably orthogonal dimensions established by factor analysis or by their own clinical intuitions. Those who believe that evaluation is either one general summary judgment or a composite of a large number

of specific observations, set up undifferentiated lists of good and bad statements, perhaps item-analyzed against some summary rating. In neither case is the rater consulted about the interpretation he would make of his own report.

This imposition of psychologists' values presupposes both understanding among psychologists concerning the organization of traits, and agreement among raters about the interpretations of various forms of behavior in relation to these traits. Without agreement, at least among a plurality of psychologists, impositions of their interpretations upon others seems presumptuous to say the least. Without consensus among the raters, more importantly, the raters cannot be expected to utilize the scales offered to them with any conviction or agreement.

This kind of consensus can be achieved only if the persons who will be rating indicate, in their own terms, what kind of behavior represents each level of each discriminably different characteristic, and which trait is illustrated by each kind of behavior.

Moreover, the rater must be "sold" upon the desirability of completing the ratings honestly and carefully, which means that the rating scales must have face validity for the purposes of the rater (which include guidance and counseling) as well as those of the researcher. Participation in the rating program must be elicited by virtue of the apparent usefulness of the procedure. These requirements are superimposed upon such

¹This research was sponsored and financed by the National League for Nursing. Thanks are extended to the numerous League staff members who contributed their time and advice. Particular gratitude is due to Phyllis Heslin who edited most of the items and was responsible throughout for organization and follow up of the data-collection procedures.

essential measurement requirements as interrater reliability and independence of scales.

The present scales were constructed to meet the needs of raters in extremely diverse situations—head nurses rating the performance of staff nurses in a variety of hospitals, under a wide range of working loads, and with a diversity of previous training of both raters and ratees. Despite the differences in raters, they can be reasonably expected to share some common core of experience and of values concerning behavior on the jobs they will rate. The situation is similar to that in most executive, administrative, and technical positions, in which jobs with a single title are seldom comparable in either level of performance required or dimensions of performance considered important. In many of these situations, moreover, participation in the rating program cannot be demanded or even “persuasively encouraged” by higher echelon personnel, but must be voluntary.

The format proposed for the rating scales is a series of continuous graphic rating scales, arranged vertically, in a manner similar to that of the Fels scales (Champney, 1941). Behavioral descriptions, exemplifying various degrees of each dimension, are printed beside the line at different heights according to their scale positions as determined by judgments of head nurses similar to those who will be expected to use the scales. The examples are intended as anchors to define levels of the characteristic, and as operational definitions of the dimension being rated. Ratings are to be made by checking at any position along the line, since summaries (see Garner, 1960) suggest that information may be gained in some circumstances even when a large number of categories or a continuous scale is used. Provision is made in the format for support of each check by notes concerning actual observed behavior.

This format was chosen as a means of combining the relevance to direct observation of critical incidents and similar techniques, with the acceptability to raters of graphic rating scales. Rating errors associated with lack of definition of either dimensions or levels militated against the use of any

of the traditionally used rating scales. Use of critical incidents, although extremely desirable because of reference to observed behavior (Flanagan, 1949), was eliminated since pretests had indicated that because of variations in the nursing situation a specific critical behavior often could not occur and hence could not serve as a basis for rating; and since most critical incidents cited tend to be too extreme for good psychometric policy which requires most accurate rating near the mean, rather than at the extremes. Use of forced choice was eliminated because of potential fakability despite format (Bass, 1957; Dicken, 1959; Gordon & Stapleton, 1956; Levonian, Comrey, Levy, & Procter, 1959; Longstaff & Jorgensen, 1953; Mais, 1951) and lack of validity in some important field tests (Kay, 1959).

The unacceptability to the rater of the forced-choice format was the most crucial deciding factor. The experience of the Army with this system led to its abandonment in 1950, since “raters . . . found it so unacceptable to rate without knowledge of the final outcome that they concentrated on finding ways to beat the system” (Rogers, 1960, p. 51). In addition, forced-choice scales and home-office-scored check lists were both rejected because, in our experience, these scales include items almost as vague as those in the traditional rating scales.

The examples we used, therefore, represent not actual observed behaviors but inferences or predictions from observations. Raters are asked to decide whether a given behavior they have observed would lead them to expect behavior like that in the description. Instead of statements such as “shows interest in patients’ description of symptoms,” the anchors consist of expectations such as “If this nurse were admitting a patient who talks rapidly and continuously of her symptoms and past medical history, could be expected to look interested and listen.” Calling for the rater to make such predictions implies that he is willing to infer from observations of behavior, that he has his own—at least implicit—belief about the intercorrelation of behaviors. The present procedure gambles that among a relatively homogeneous group of judges such as head

nurses, these beliefs will be reasonably well standardized. It demands that such predictions be organized into areas not by theoretical similarity, but by judged similarity as indicated by the raters, and that the areas represent dimensions meaningful to the raters. It also provides checks in that each rater records briefly the behavior on which his prediction was based.

The use of an open, obviously fakable format assumes that raters will, under proper circumstances, give conscientious estimates of the level of performance of ratees.

We believe that most rating errors are not due to deliberate faking. Moreover, no rating scale is really proof against distortion by a rater who really wants to do so. Better ratings can be obtained, in our opinion, not by trying to trick the rater (as in forced-choice scales) but by helping him to rate. We should ask him questions which he can honestly answer about behaviors which he can observe. We should reassure him that his answers will not be misinterpreted, and we should provide a basis by which he and others can check his answers.

The use of expected behaviors is intended to encourage such conscientiousness by making the predictions (*a*) so concrete that, in view of previous agreement by the peer (head nurse) group, central tendency or hedging effects will be minimized; and (*b*) so verifiable that the insight, judgment, values, etc., of the rater are potentially challenged if later behavior of the ratee should fail to confirm the prediction.

The basic procedure for scale construction resembles that employed to ensure that

translations from one language to another adhere to the connotations as well as to the denotations of the original. Material is translated into a foreign language, and then, by an independent translator, retranslated into the original. Where "slippage" occurs, translations are corrected. Similarly, we required that examples, or expectations, be classified as indicative of a given dimension of nursing performance, and that independent judges indicate what dimension is illustrated by each. In addition, we defined the dimensions in the judges' own terminology and scaled the examples along these dimensions.

The submission of examples and subsequent reallocation by the raters' peers seems to ensure a high degree of content validity for the items and the scales.

PROCEDURES

Four groups of head nurses were sent by their hospitals to participate in conferences concerned with the use of evaluation in improving nursing performance. Data were gathered by mail from the remaining head nurses from the same hospitals from which two of the conference groups were drawn. Table 1 shows the samples used. Sample F was held out for an independent replication. The samples are diversified and probably representative of head nurses who would be likely to use such a rating instrument.

The procedure used was essentially an iterative one, work performed by one group being checked and revised by others, so that the number of judges differs for different parts of the data. The content area was restricted to that of medical-surgical nursing.

1. First, qualities or characteristics to be evaluated were listed by each group; the most frequent dimensions were selected for further analysis. The nurses' own terminology was retained. Coverage

TABLE 1
JUDGES USED IN SUCCESSIVE SAMPLES

Sample	<i>N</i>	Situation	Geographical distribution of judges
A ^a	85	Conference	New York City and Northeastern States
B	154	Mailing	Same hospitals as Sample A
C	88	Conference	Entire continental United States with concentration in Ohio and Kentucky
D	130	Mailing	Same hospitals as Sample C
E	85	Conference	Midwest, primarily
F	83	Conference	New England

Note.—All judgments reported were made individually and reported independently, whether gathered by mail or in conferences.
^a Revision of general standards and item writing only, by this sample.

of important aspects was further insured by gathering and classifying critical incidents in the customary way, for Sample A.

2. The groups formulated general statements representing definitions of high, low, and acceptable performance for each quality.

3. The groups submitted examples of behavior in each quality, and these were edited into the form of expectations of specific behavior.

4. Judges indicated, independently, what quality was illustrated by each example. *Examples* were eliminated if there was not clear modal agreement as to the quality to which each belonged. *Qualities* were eliminated if examples were not consistently reassigned to the quality for which they were originally designed.

5. Other judges used the examples to describe a specific nurse with outstandingly good nursing performance and another nurse with unsatisfactory performance. The difference between the outstanding and unsatisfactory nurse was computed for each pair of ratings to determine the discrimination value for each example.

6. Each vertical scale, together with the general definitions, was presented with a list of items previously judged by other raters as belonging to that quality. Judges rated each item from .0 to 2.0 according to the desirability of the behavior illustrated. Items were eliminated if the dispersion of judgments was large, or if the distribution was multimodal. All of the items which met these criteria were assembled for each scale, and the mean scale positions assigned to them for each group of judges were intercorrelated to give estimates of scale reliabilities. As indicated by Noble (1955) "if one is interested in the consistency with which successive random samples of subjects or judges respond to an invariant set of stimulus items, one encounters a problem of *scale reliability* rather than one of *test reliability*" (p. 195). It was recognized that mean values were distorted somewhat by skewness in the distributions of judgments at the extremes, but it was felt that since the effect of skewness would be to reduce the stability of the means as estimates of central tendency, correlations using means would give at least a minimal estimate of agreement among groups of judges as to the relative position of items on a scale.

In addition, a comparison among samples of means and variances for all items in each scale indicates the similarity in the absolute location of the items by the various groups of judges.

RESULTS

The qualities which were most frequently considered important are listed in the first column of Table 2. An attempt was made to construct a scale and it proved possible to write general definitions about which there was considerable agreement in each of these areas. Examples of expected behavior were

TABLE 2
AGREEMENT CONCERNING ALLOCATION OF
ITEMS TO CHARACTERISTICS

Characteristic	High agreement (above 59%)	Lower agreement (40%-59%)
Knowledge and Judgment	16	7
Conscientiousness	11	8
Skill in Human Relationships	21	9
Organizational Ability	16	4
Communication Skills	1	4
Objectivity	2	8
Flexibility	5	1
Reaction under Pressure	1	1
Observational Ability	7	4
Total number of items	80	46
Percentage of 141 items presented	57%	33%

Note.—Samples C and E combined.

submitted which met the standards of group agreement for each.

The retranslation procedure, however, eliminated many items and several qualities. Table 2 shows the number of items surviving for each. Some of the eliminations are interesting in themselves; items designed to illustrate Reaction under Pressure, for example, were frequently allocated to Organizational Ability or Knowledge and Judgment, on the grounds that a certain degree of crisis is normal in nursing and ability to meet it involves primarily establishing priorities and knowing what to do. Communication Skills items were allocated to Skill in Human Relationships or Conscientiousness because the items involved, to a large extent, either explaining to patients or keeping records. Both Reaction under

TABLE 3
AGREEMENT OF SAMPLE F WITH COMBINED SAMPLES
B TO E ON ASSIGNMENT TO MODAL QUALITY

Quality	Number of items assigned by both groups	Geisser's Concomitance Measure R
Knowledge and Judgment	16	.52*
Conscientiousness	12	.55*
Skill in Human Relationships	23	.84*
Organizational Ability	18	.90*
Objectivity	9	.71*
Observational Ability	11	.83*

* $p \leq .01$.

TABLE 4
INTERCORRELATIONS AND VARIANCES OF MEAN SCALE POSITIONS ASSIGNED BY FOUR
SAMPLES FOR THE FOUR SCALES WITH LARGEST NUMBERS OF ITEMS

Characteristic scaled	Number of items	Sample of judges ^a	Correlations			Mean	σ^2
			C	D	E		
Knowledge and Judgment	18	B	.983	.972	.979	.999	.210
		C	—	.986	.981	.897	.189
		D		—	.980	.983	.175
		E			—	.985	.174
Conscientiousness	14	B	.991	.996	.995	.817	.304
		C	—	.996	.995	.869	.353
		D		—	.995	.801	.225
		E			—	.786	.220
Skill in Human Relationships	20	B	.987	.992	.986	.954	.221
		C	—	.992	.991	.896	.219
				—	.991	.928	.183
					—	.847	.176
Organizational Ability	17	B	.990	.990	.992	.934	.306
		C	—	.992	.987	.874	.222
		D		—	.991	.911	.208
					—	.983	.273

Note.—The five additional scales contained too few items for correlational purposes and were arbitrarily grouped for analysis. Correlations for these groups ranged from .994 to .997. The items included only those originally designed for each scale.

^a For sample characteristics, see Table 1.

Pressure and Communication Skills were eliminated at this step. Flexibility was also considered unpromising, although attempts were made to retrieve it in later institutes not reported here.

The significance of agreement of judges in assigning items to the same modal classification was tested for six scales (see Table 3) by a test of concomitance (Geisser, 1958) for all items which, after editing, had been presented for allocation to areas or qualities. This coefficient showed agreement to be significantly above chance when the judgments of Sample F were compared with judgments of previous groups (Table 3). It should be noted that this agreement was tested for examples presented in a scrambled order and judged one at a time; much better agreement can be expected when several examples of the same quality are grouped together on a single page as in the proposed rating scale format. A separate test of allocation of a few pairs and triads of items showed very high agreement.

Discrimination of items was checked by

comparing average ratings assigned to outstanding and unsatisfactory nurses. For all items passing the previous criteria the differences were significant by the chi square test. Also, virtually all differences for individual judges were positive. Therefore all retained items were clearly relevant in discriminating extreme levels of performance.

TABLE 5
CROSS-CHECK OF SCALE RELIABILITIES

Characteristic scaled	Number of items	<i>r</i>
Knowledge and Judgment	21 ^a	.987
Conscientiousness	16 ^a	.997
Skill in Human Relationships	19 ^a	.993
Organizational Ability	20 ^a	.990
Objectivity	10 ^b	.998
Observational Ability	10 ^b	.982

Note.—Correlations of mean scale positions (457 judges) with mean values for hold-out group (83 judges).

^a Scales included several items originally scaled for other characteristics and reallocated for the analysis of the hold-out sample.

^b Scales included tentatively. Some items judged by only Samples C and E (*N* = 173).

Scale reliabilities for the first four samples of judges for all retained items are given in Table 4. The mean evaluation of each of these items for four samples was correlated with that given to it by the holdout sample (F) and is given in Table 5. The lowest scale reliability is .972. Grand means and variances for all items in each scale are given for each sample in Table 4. Differences among the grand means within each scale were not significant by F tests.

Parenthetically, there are no consistent or significant differences in intercorrelations, means, or variances between Conference and Mail groups.

DISCUSSION

In general, the procedure seems satisfactory, with adequate agreement concerning allocation of examples, excellent discrimination, and high scale reliability. The potential advantages of scales based on such procedures are obvious; they are rooted in, and referable to, actual observed behavior; evaluations of the behavior have been made by judges at least reasonably comparable to those who will eventually use the scales; whatever we think of the terminology, the traits or qualities covered are operationally defined and are distinguishable one from another by the raters. Both dimension and level have been agreed upon, so that there is a fair chance of treating ratings by different raters as comparable just so long as they agree with the interpretations of the expectations. Even though different specific behaviors may be observed in different situations, they are referred to the common set of expectations which serves as a mutual frame of reference. Moreover, the chance to supplement predictions with documentation of the actual observed behavior upon which the predictions are based (as is provided by the scale format) permits checking and revision of examples and scales. Also, the use of these supporting anecdotal records, as well as the ease with which predictions can later be checked, favors honest and conscientious rating.

The disadvantages are equally obvious. The decision to use the raters' own theory of traits, and retranslation as a method of checking items, implied that raters would

eventually be able to make decisions concerning what characteristics were involved in a given observed behavior. Most behavior is complex, and certainly not attributable to a single cause in the makeup of the individual, without regard to interaction of needs or to the influence of the situation. The use of trait names, and of general statements concerning levels of performance, in addition to the behavioral anchors may make for ambiguity in ratings, especially if one set of raters is less critical than another and displaces the items in relation to the general standards. One set of raters may also be more likely to attribute complex behavior to one cause while another set may prefer another. There are also too many scales for easy handling.

We hope that the number of qualities can be reduced after actual field use permits computation of scale intercorrelations and interrater reliabilities under normal rating conditions. We expect that experience and training with the scales will enable the rater to evaluate complex items of behavior on several relevant scales (and that this evaluation may serve some useful function in improving the rater's ability to interpret, diagnose, and improve the behavior of the person rated). We further hope that the general trait names, with their accompanying general definitions, will "wither away" and be replaced in use by the operational definitions provided by the behavioral expectations. Grouping the expectations on a single page will improve agreement as to qualities also. We expect, moreover, that as higher standards of performance become generally accepted, examples may appear that may be placed even higher in the vertical scales than the present items.

Weights for combining scales in order to give a summary rating, where needed, should be determined empirically by the use of multiple regression against a reliable rating of overall performance obtained in the field and probably separately for each nursing situation.

The consistency of judgments renders tenable the hypothesis that homogeneous groups of judges do share a common belief about the manner in which behaviors are

intercorrelated and can extrapolate from observed to predicted behavior.

The procedure seems promising not only for nursing, but also for other complex tasks. Parenthetically, we should point out that reliabilities are so high that procedures similar to this one could certainly be attempted with smaller numbers of judges, and that sampling differences seem to be a relatively trivial source of error. Wherever behaviors may be expected to be reasonably comparable or interpretable from one situation to another, as in many professional and administrative jobs, and in research settings where observations can be made under fairly uniform conditions, the procedure seems applicable. We hope that it will prove useful in industrial, educational, and social areas of research.

REFERENCES

- BASS, M. Faking by sales applicants of a forced choice personality inventory. *J. appl. Psychol.*, 1957, **44**, 403-404.
- CHAMPNEY, H. The measurement of parent behavior. *Child Develpm.*, 1941, **12**, 131-166.
- DICKEN, C. F. Simulated patterns on the Edwards Personal Preference Schedule. *J. appl. Psychol.*, 1959, **43**, 372-378.
- FLANAGAN, J. C. A new approach to evaluating personnel. *Personnel*, 1949, **26**, 35-42.
- GARNER, W. R. Rating scales, discriminability, and information transmission. *Psychol. Rev.*, 1960, **67**, 343-352.
- GEISSER, S. A note on McQuitty's index of concomitance. *Educ. psychol. Measmt.*, 1958, **18**, 125-128.
- GORDON, L. V., & STAPLETON, E. S. Fakability of a forced-choice personality test under realistic high school employment conditions. *J. appl. Psychol.*, 1956, **40**, 258-262.
- KAY, B. R. The use of critical incidents in a forced-choice scale. *J. appl. Psychol.*, 1959, **43**, 269-270.
- LEVONIAN, E., COMREY, A., LEVY, W., & PROCTER, D. A statistical evaluation of Edwards Personal Preference Schedule. *J. appl. Psychol.*, 1959, **43**, 355-359.
- LONGSTAFF, H. P., & JURGENSEN, C. E. Fakability of the Jurgensen Classification Inventory. *J. appl. Psychol.*, 1953, **37**, 86-89.
- MAIS, R. D. Fakability of the classification inventory scored for self-confidence. *J. appl. Psychol.*, 1951, **35**, 172-174.
- NOBLE, C. E. Scale reliability and the Spearman-Brown equation. *Educ. psychol. Measmt.*, 1955, **15**, 195-205.
- ROGERS, B. F. The current status of the United States Air Force Officer Effectiveness Report. Unpublished master's thesis, Florida State University, 1960.

(Received May 28, 1962)

SELF-CONFIDENCE AS A RESPONSE SET¹

CECIL J. MULLINS

Air Force Systems Command, Lackland Air Force Base, Texas

2 tests, 1 spatial, the other verbal, were designed so that half the items on each test did not contain a correct alternative. Every item had "Correct answer not given" as one of the alternatives. The responses to the alternative Correct answer not given correlate across the tests, even when the abilities measured by the 2 tests are partialled out. The Correct answer not given alternative is relatively independent of the aptitudes measured by the 2 tests.

If a test were so designed that each item of it contained alternatives chosen deliberately so that they ranged from completely wrong to almost right, and if one of the alternatives for each item were "Correct answer not given," it seems reasonable that one might find a response set such that some people would consistently tend to check the Correct answer not given response whereas others might consistently avoid it.

Obviously, the test would have to be disguised in some way, since some of the subjects would soon notice that Correct answer not given is always the correct response. One way to handle the problem is to design a test each item of which has the Correct answer not given alternative, but which contains some items which have the correct answers listed among the alternatives, intermingled with other items in which the Correct answer not given alternative is the correct response. If we mixed these two types of items equally, we would also have a built-in device for determining whether or not the Correct answer not given alternative is measuring anything other than ability on the test. We could consider the half of the test which contains the correct answer among the given alternatives as a test of ability and the other half of the test as a measure combining ability with the response set we are seeking. If we scored only the Correct answer not given response, whether it appeared in the half of the test in which the correct answer is given or in the other half of the test, we would probably have a measure which is

more purely a measure of the response set. For convenience of expression, we shall call the Correct answer not given alternative the S response (for self-confidence), since the author believes that self-confidence is the personality characteristic which determines this set.

Furthermore, if we designed two such tests in two different formats, such as spatial reasoning items and vocabulary items, and if we expect the S response to measure a generalized test set, we should reasonably expect the S response score obtained on the vocabulary test to correlate with the S response obtained on the spatial reasoning test.

TESTS AND SUBJECTS

Two tests were constructed along the lines indicated above. One was called Spatial Reasoning-A, and the other was Vocabulary-C. Spatial Reasoning-A contains 46 items, all of which call for spatial serial reasoning. Each item begins with a series of four diagrams which change in some logical fashion from the first through the fourth. The subject is to discover the logic in the first four diagrams, decide what the next diagram would look like if the logic were continued one more step, and pick out of five more diagrams the one which is correct. The five offered diagrams constitute five alternatives for each item, and each item contains one extra alternative which is to be marked if the subject believes the correct diagram is not among the five offered. Twenty of the items have the correct answer presented among the offered alternatives. All the others have at least one nearly correct alternative among those offered.

Vocabulary-C is a test of 40 vocabulary items. Each item has a Correct answer not given alternative along with four other alternatives. Half of the items have the correct answer given; the other half do not. Where the correct answer is not given, an effort was made to provide at least one wrong answer which was almost (but not quite) right.

Both these tests were administered to nine flights

¹ The research reported in this paper was sponsored by the 6570th Personnel Research Laboratory, Aerospace Medical Division, under AFSC Project 7717(05).

TABLE 1

INTERCORRELATIONS AMONG SIX SCORES FROM VOCABULARY-C AND SPATIAL REASONING-A

	2	3	4	5	6	<i>M</i>	<i>SD</i>	<i>r_{tt}</i>
1. Vocabulary-C Ability	.25	-.01	.38	.34	.22	9.76	4.45	.80
2. Vocabulary-C, Right S		.89	.21	.41	.42	8.78	3.95	.81
3. Vocabulary-C, Total S			.10	.36	.38	12.56	5.73	.60
4. Spatial Reasoning-A, Ability				.55	.31	10.57	4.03	.68
5. Spatial Reasoning-A, Right S					.92	11.71	5.66	.85
6. Spatial Reasoning-A, Total S						15.46	6.74	.79

Note.—*N* = 436. $r_{25,14} = .34$, $r_{36,14} = .38$. .10 significant at .05 level, .13 significant at .01 level.

of basic airmen (*N* = 436), along with several other tests. Intercorrelations were computed on the following six scores derived from the two tests:

Vocabulary-C, Ability. The total rights score on the half of the test in which the correct answer was listed among the alternatives.

Vocabulary-C, Right S. The total rights score on the half of the test in which the correct answer was not listed among the alternatives. In other words, this score represents the number of times the S response was chosen when the S response was correct.

Vocabulary-C, Total S. The total number of S responses chosen, regardless of whether the S response was the correct one or not.

Spatial Reasoning-A, Ability. The total rights score on the part of the test in which the correct answer was listed among the alternatives.

Spatial Reasoning-A, Right S. The total rights score on the part of the test in which the correct answer was not listed among the alternatives.

Spatial Reasoning-A, Total S. The total number of S responses chosen, regardless of whether the S response was the correct one or not.

RESULTS AND DISCUSSION

The intercorrelations among these six scores are given in Table 1. The reliabilities are odds-evens, corrected by the Spearman-Brown prophecy formula. It should be noted that the correlation between rights scores on the section of the test in which the correct answer is among those presented as choices to the subject and those on the section of the

test in which the correct answer is not among those given is only .25 for the Vocabulary-C test and .55 for the Spatial Reasoning-A test. Since those correlations are considerably below the reliabilities of the scores correlated, it appears that something is being measured by the S response not contained in the ability score.

The ability scores correlate even less with the Total S scores (−.01) for Vocabulary-C and .31 for Spatial Reasoning-A). The Total S score obtained from Vocabulary-C correlates with the Spatial Reasoning-A, Total S score .38. Right S correlates across the two tests .41.

The ability measures across the two tests correlate .38, however, which confuses the picture somewhat. The next step was to compute a partial correlation coefficient between Rights S scores across the two tests, with vocabulary ability and spatial reasoning ability both held constant. This was done, and the partial *r* resulting was .34.

As a last step, another partial *r* was computed between the Total S scores across the two tests, holding vocabulary and spatial reasoning ability constant. The partial *r* emerging from this computation was .38.

(Received June 8, 1962)

TWO APPROACHES TO THE PREDICTION OF GROUP RESPONSES

ROSSALL J. JOHNSON

School of Business, Northwestern University

The method of measuring the ability of an individual to predict the responses of a group was examined. A comparison was made of a predictor's score when summing his correct individual predictions and the score obtained by summing his correct prediction in one direction as compared to the number of responses in that direction. A relatively low correlation was obtained.

This paper is concerned with the method of measuring the ability of the individual to predict the responses of a group.

In previous studies (Anikeeff, 1951; Nagle, 1953; Patton, 1951; Singh, 1953; Tagiuri, 1958; Tannenbaum, Weschler, & Massarik, 1961) individuals have attempted to predict how certain groups of people would answer specific questions. Individuals in one ethnic group would predict the percentage of persons in another ethnic group that would give a specific response to a particular question or a supervisor would predict the percentage of subordinates who would answer questions in a specified direction.

The difficulty encountered in this method of measurement can be demonstrated by the following example:

A supervisor predicts that 50% of his subordinates will answer YES to the question: "Do you feel you have a good future with the company?" After tabulating the results, 50% did answer YES. From this a conclusion may be drawn that the supervisor "knows" his men. However, the supervisor could be 100% wrong if the 50% *he was thinking about* answered it NO and the other 50% answered YES.

Many artifacts seem to be present when predicting responses of a group as a whole rather than considering the responses of the individuals within the group. For this reason the question of measurement of individual responses within a group arose.

In order to determine experimentally if this was a problem in the measurement of predicting ability, the following null hypothesis was set up:

There is no significant relationship between the percentage of correct subordinate responses predicted when calculated by the supervisor group scoring method and the percentage of correct subordinate responses predicted when calculated by the supervisor individual scoring method.

PROCEDURE

A sample of 227 subordinates and 25 supervisors was taken from two companies. The subordinate, for the purpose of this study, is designated as a randomly selected hourly paid worker who does not have group leader responsibility and who has worked for the tested supervisor for at least 9 months. The supervisor is defined as a salaried supervisor who has at least 12 subordinates (as defined above) reporting directly to him. Eight to 10 subordinates under each of the 25 supervisors participated in the project.

The subordinate questionnaire consisted of 40 questions. Twenty questions were selected from Form A of the test *How Supervise?* These questions were from the sections on supervisory practices and supervisory opinions. The question mark or undecided response was omitted.

The other 20 questions were morale questions. In a previous study (Harris, 1949), these questions had D values (Lawshe, 1942) of 1.10 or higher.

Each subordinate was guaranteed anonymity. His name and personal data were on a separate sheet deposited in a ballot type of box while the questionnaire with the supervisor's name only was deposited in another box. This questionnaire and personal data sheet could later be brought back together by means of a code.

The supervisor questionnaire contained the same questions as the subordinate questionnaire but with different instructions. The supervisor was also given a list of the names of 10 of his subordinates and a matching code number for each subordinate; e.g., Bill Jones, 1; Irving Smith, 2; Jack Brown, 3; etc. The instructions to the supervisor were as follows:

The following questions have been answered by the workers that you supervise. How well do you

know your subordinates? Can you guess how they answered these questions?

You have been given a list of your subordinates with identifying numbers. These subordinates have answered the following questions. Mark on the appropriate line the identifying numbers of the subordinates that you think gave that response.

Here is an example of how we'd like you to guess how they answered these questions:

Do you enjoy going to a ball game on Sunday?
 Employees 2, 3, 7, 9, 10 probably said LIKE
 Employees 1, 4, 5 probably said DIS-LIKE
 Employees 6, 8 probably said IN-DIFFERENT

You have 10 subordinates on the list. If you think that individuals 2, 3, 7, 9, and 10 like going to a ball game on Sunday, then mark these numbers in the space above so that the sentence reads, "Employees 2, 3, 7, 9, and 10 said LIKE." If you think that individuals 1, 4, and 5 dislike going to a ball game on Sunday, then write the numbers 1, 4, and 5 in the space for the word "DISLIKE." If you feel that subordinates 6 and 8 neither like nor dislike going to the ball game, that is, they are indifferent, then write the numbers 6 and 8 in the appropriate space (as shown above).

The supervisor questionnaire was scored in two ways:

*Group Scoring:*¹ The number of times the subordinates answered a given question was matched with the number of times the supervisor predicted the group would give that response. The individual responses were counted only as part of the group and not matched with the supervisor on an individual basis. Thus, if out of 10 subordinates 6 gave the answer "LIKE" and the supervisor predicted that 4 individuals gave the answer "LIKE" then his score for that question was 4 regardless of which 4 he predicted would give that response. His total score was the percentage of the total that was correctly predicted.

*Individual Scoring:*² The score is the percentage of times the supervisor correctly predicts the responses of those subordinates involved. For instance, if he correctly predicts 210 times of the total of 400 times that 10 subordinates answered questions, then his score would be 52.2%.

The prediction scores as calculated by the supervisor group scoring method were matched with prediction scores as calculated by the supervisor individual scoring method. Pearson's coefficient of correlation between these two scores was calculated

separately for the morale questions and the How Supervise? questions.

RESULTS

The hypothesis is rejected. When based upon the 20 Form A How Supervise? questions, the Pearson coefficient of correlation between the supervisor group scoring method and the supervisor individual scoring method was +.55.

When the 20 morale survey questions were used, the Pearson coefficient of correlation between supervisor group scoring method and supervisor individual group scoring method was +.57.

Both correlations were significant at the 1% level.

CONCLUSIONS AND DISCUSSION

Although there is a positive correlation between the supervisor predicting scores as calculated by the group method and by the individual method, this relationship is not large enough to justify using the scores interchangeably. More specifically, it would not seem advisable to use the group scoring method as a reasonable estimate of the individual score. The supervisor's predicting ability is more accurately portrayed when the prediction can be compared with the individual responses rather than compare the sum of the individual predictions with the sum of the responses.

If one is of the opinion that the sum of the individual responses represents the group response, then the above procedure of comparing individual prediction with individual response is superior. However, it has been argued that the sum of the individual responses may not be the same thing as group response. It may be that if one could record the responses of the group as a unit or whole it would be different from the sum of individual responses. This would be true just as group action is different from the sums of the individual actions.

Does the supervisor make a decision based on the so-called "group as a whole" response, or does he base it upon the "sum of the individual" responses? For instance, if the supervisor is trying to make a decision on whether or not to work the group overtime

¹ The penalty for not predicting exactly is that the supervisor loses if he overestimates or underestimates.

² The supervisor is penalized if he does not predict the exact responses of an individual.

on a certain night, does he think in terms of individuals? Does he say to himself: "Bill and Bob will sure object but Art, Ralph, Joe, and the rest of the fellows will be for it"? Or will he say to himself: "Most of the gang wants to work overtime"?

When the superintendent asks the supervisor, "How do you think the fellows will feel about the wage increase?" does the supervisor reply: "Oh, I think all except one or two will like it," or does he say: "Everyone will like it except Phil and Sam and they will probably want more"?

These examples of group predicting versus individual predicting are true examples, only if, when the supervisor says, "Most of the gang," he is not thinking specifically of Bill and Bob. Again, when he says, "Except one or two" and is not thinking specifically of Phil and Sam, only then could it be considered group predicting.

It may be that one could discover whether or not the supervisor was thinking about Bill and Bob or Phil and Sam. If he named them in response to an inquiry, then you would say he was thinking in terms of individuals. If, on the other hand, he said: "Oh, there are always one or two who won't go along with the group," then he is thinking in terms of the group as a whole.

It is to be expected that there would be some relationship between the two scoring methods since several artifacts are involved and probably additional ones that are not immediately obvious. Certainly the usual contamination from projection is involved in

both scoring methods, but, by using the individual method, the supervisor projecting most will receive the lowest score *unless* all of the subordinates answer the questions the same way and in the direction the supervisor projects.

This problem of group versus individual response cannot be resolved in this study. It can, however, be pointed out that the individual responses in this study, when summed, did not accurately portray a group-as-a-whole response.

REFERENCES

- ANIKEEFF, A. M. Reciprocal empathy: Mutual understanding among conflict groups. *Stud. higher Educ.*, 1951, 77, 1-43.
- HARRIS, F. J. The quantification of an industrial employee survey: I. Method. *J. appl. Psychol.*, 1949, 33, 103-111.
- LAWSHE, C. H., JR. A nomograph for estimating the validity of test items. *J. appl. Psychol.*, 1942, 26, 846-849.
- NAGLE, B. F. Productivity, employee attitude and supervisor sensitivity. Unpublished doctor's dissertation, Purdue University, 1953.
- PATTON, W. M. A study of certain psychological variables related to supervision in the textile industry. Unpublished doctor's dissertation, Purdue University, 1951.
- SINGH, SITARAM. An experimental study of reciprocal empathy of executives and their subordinates. Unpublished doctor's dissertation, Purdue University, 1953.
- TAGIURI, R., & PETRULLO, P. *Person perception and interpersonal behavior*. Stanford: Stanford Univer. Press, 1958.
- TANNENBAUM, R., WECHSLER, I., & MASSARIK, F. *Leadership and organization*. New York: McGraw-Hill, 1961.

(Received June 18, 1962)

Journal of Applied Psychology

KENNETH E. CLARK, Editor
University of Colorado

MARYGROVE COLLEGE LIBRARY
DETROIT, MICHIGAN
PLEASE DO NOT REMOVE

Table of Contents

Allocation of Functions between Man and Machines in Automated Systems: Nehemiah Jordan	161
A Factor Analytic Study of Perceived Occupational Similarity: George G. Gonyea and Clifford E. Lunneborg	166
Seniority in Work Groups: A Right or an Honor?: Norman R. F. Maier and L. Richard Hoffman	173
Auditory Perception of Position and Speed: Leslie Buck	177
Creativity Tests and Achievement in High School Science: Victor B. Cline, James M. Richards, Jr., and Walter E. Needham	184
Effects of Magnification on a Subminiature Assembly Operation: Ravi M. Nayyar and J. Richard Simon	190
Effect of Reference Marks on the Detection of Signals on a Clock Face: Jane F. Mackworth	196
Series Effects in Motor Performance Studies: J. E. Kennedy and J. Landesman	202
Assessments of Innovative Behavior: Partial Criteria for the Assessment of Executive Performance: Garlie A. Forehand	206
Knowledge of Results and Signal Rate in Monitoring: A Transfer of Training Approach: Earl L. Wiener	214
A Note on Job Performance: Differences between Respondent and Nonrespondent Salesmen to an Attitude Survey: Wayne K. Kirchner and Nancy B. Mousley	223

American Psychological Association

Consulting Editors

GEORGE E. BRIGGS, *Ohio State University*
NORMAN FREDERIKSEN, *Educational Testing Service*
LEONARD D. GOODSTEIN, *University of Iowa*
EDWIN R. HENRY, *Standard Oil Company of New Jersey*
JOHN HOLLAND, *National Merit Scholarship Corporation*
CLIFFORD E. JURGENSEN, *Minneapolis Gas Company*

LAURENCE S. MCGAUGHRAN, *University of Houston*
QUINN MCNEMAR, *Stanford University*
HAROLD F. ROTHE, *Beloit Corporation*
THOMAS A. RYAN, *Cornell University*
ALEXANDER G. WESMAN, *Psychological Corporation*
CLARK L. WILSON, *Batten, Barton, Durstine, and Osborn, Incorporated*

This journal gives primary consideration to original investigations in any field of applied psychology except clinical psychology, although a descriptive or theoretical article may be accepted if it represents a special contribution in an applied field. Quantitative investigations of interest or value to psychologists working in the following broad fields will be considered: vocational and educational prognosis, diagnosis, and guidance at the secondary and college level; personnel research in business, industry, and government; engineering psychology; industrial working conditions; research on opinion and morale factors; job analysis and classification research; market and advertising research.

Manuscripts must be accompanied by an abstract of 100–120 words typed on a separate sheet of paper. The abstract should conform to the style of *Psychological Abstracts*. Detailed instructions for preparation of the abstracts appeared in the *American Psychologist* (1961, 16, 833), or they may be obtained from the Editor or from the APA Central Office.

Manuscripts should be addressed to the Editor:

Dr. Kenneth E. Clark
College of Arts and Sciences
302 Morey Hall
University of Rochester
Rochester 20, New York

All manuscripts must be submitted in duplicate. Original figures are prepared for publication; duplicate figures may be photographs or pencil-drawn copies.

Manuscripts must conform to the style requirements described in the *Publication Manual of the American Psychological Association*.

The following policies govern reprinting of materials copyrighted in APA journals: (a) to require approval to reprint for tables and figures, and for text only if more than 500 words [total from one article] in length; (b) to continue approval to reprint, contingent on the author's approval, for articles reprinted in whole or in major part; (c) to negotiate where possible the dedication of royalties from commercial publishers to the American Psychological Foundation.

Journal of Applied Psychology

Published bimonthly by the
American Psychological Association
Prince and Lemon Sts., Lancaster, Pa.
and 1333 Sixteenth Street N. W.
Washington 6, D. C.

\$10.00 per volume

\$2.00 per issue

HELEN ORR
Managing Editor

ELIZABETH S. REED
Advertising Manager

VIRGINIA RICHARDS
Editorial Assistant

Subscriptions, orders, and business communications should be addressed to the American Psychological Association, 1333 Sixteenth St. N.W., Washington 6, D. C. Address changes must reach the subscription office by the tenth of the month to take effect the following month. Undelivered copies resulting from address changes will not be replaced; subscribers should notify the post office that they will guarantee second-class forwarding postage. Other claims for undelivered copies must be made within four months of publication.

Second class postage paid at Lancaster, Pennsylvania and at additional mailing places.

© 1963 by the American Psychological Association, Inc.

Journal of Applied Psychology

VOL. 47, No. 3

JUNE 1963

ALLOCATION OF FUNCTIONS BETWEEN MAN AND MACHINES IN AUTOMATED SYSTEMS

NEHEMIAH JORDAN

RAND Corporation

With the growing complexity of the man-machine systems the problem of allocation becomes more critical. Little progress has been made towards its solution since the publication of Fitts' article in 1951 which has dominated thinking in this area. Fitts recommended that man be compared to machines and be chosen for those functions which he does better than machines and vice versa. To do so is wrong; when we can compare a man to a machine, we find that we can also build a machine for the function involved. Hence the lack of progress. Men and machines are complementary, rather than comparable. Once the problem is so reformulated, new ways of thinking which appear to be promising open up.

In a document entitled *Factors Affecting Degree of Automation in Test and Checkout Equipment* which, among other things, reviews the problems of allocation of functions, Swain and Wohl (1961) assert:

A rather stark conclusion emerges: *There is no adequate systematic methodology in existence for allocating functions* (in this case, test and checkout functions) *between man and machine*. This lack, in fact, is probably the central problem in human factors engineering today. . . . It is interesting to note that ten years of research and applications experience have failed to bring us closer to our goal than did the landmark article by Fitts in 1951 (p. 9).

These two competent and experienced observers summarize 10 years of hard and intensive labor as having basically failed. This is a serious problem. Why this failure?

We can attempt to seek a possible answer to the question by seeking a similar case in other fields of scientific endeavor and seeing what can be learned from it. And another case is easy to find; it is in fact a classical case. In their book, *The Evolution of Physics*, Einstein and Infeld (1942) spend some time discussing the problems which beset prerelativity physics in which they focus upon the concept of "ether." They point out

that ether played a central role in physical thinking for over a century after having first been introduced as a necessary medium for propagating electromagnetic waves. But during all this time all attempts to build and expand upon this concept led to difficulties and contradictions. A century of research on ether turned out to be sterile in that no significant advance was made during that time. They conclude: "After such bad experiences, this is the moment to forget ether completely and try never to mention its name" (p. 184). And they do not mention the concept anymore in the book. The facts underlying the concept were not rejected, however, and it was by focusing upon the *facts* while rejecting the *concept* that Einstein could solve the problems which bedeviled the physics of his day.

The lesson to be learned from this momentous episode is that when a scientific discipline finds itself in a dead end, despite hard and diligent work, the dead end should probably not be attributed to a lack of knowledge of facts, but to the use of faulty concepts which do not enable the discipline to order the facts properly. The failure of human factor engineering to advance in the area of allocation of functions seems to be

such a situation. Hence, in order to find an answer to the question, "Why this failure?", it may be fruitful to examine the conceptual underpinnings of our contemporary attempts at allocating functions between men and machines. And this brings us back to the landmark article by Fitts (1951) mentioned earlier.

This article gave rise to what is now informally called the "Fitts list." This is a two-column list, one column headed by the word "man" and the other, by the word "machine." It *compares* the functions for which man is superior to machines to the functions for which the machine is superior to man. Theoretically, this leads to an elegant solution to the allocation of functions. Given a complex man-machine system, identify the functions of the system and then, based on such a list which was expected to be refined with time and experience, choose machines for the functions they are best suited for and men for the functions they are best suited for. This is a clean engineering approach and it is not surprising that great hopes were placed upon it, in 1951. The only gimmick is that it did not and does not work.

The facts to be found in all the existing versions of the Fitts list are all correct, just as the facts underlying the concept ether were all correct. Hence the inutility of these lists must be attributed to what we are told to do with these facts, to the instruction to compare man to the machine and choose the one who fits a function best. I question the *comparability* of men and machines. If men and machines are not comparable, then it is not surprising that we get nowhere when we try to compare them. Just as the concept of ether led to inutility, perhaps the concept of man-machine comparability does the same. Let us explore somewhat the background to the concept *comparability*.

The literature on the place of a man in man-machine systems converges to two posthumous articles by K. J. W. Craik (1947a, 1947b). These articles are recognized by almost all as being the basis upon which much that followed is built. Craik argues that in order to best be able to plan, design, and operate a complex system man functions and

machine functions should be described in the same concepts, and, by the very nature of the case, these concepts have to be engineering terms. In other words, Craik recommends that we describe human functions in mathematical terms *comparable* to the terms used in describing mechanical functions.

In fairness to Craik's memory it must be stressed that these two papers published after his death were notes for a discussion and probably not meant for publication. Hence he should not be blamed for failing to recognize the simple fact that anytime we can reduce a human function to a mathematical formula we can generally build a machine that can do it more efficiently than a man. In other words, to the extent that man becomes comparable to a machine we do not really need him any more since he can be replaced by a machine. This necessary consequence was actually reached but not recognized in a later paper, also a fundamental and significant paper in human factor engineering literature. Birmingham and Taylor (1954) in their paper, "A Design Philosophy for Man-Machine Control Systems," write: "speaking mathematically, he (man) is best when doing least" (p. 1752). The conclusion is inescapable—design the man out of the system. If he does best when he does least, the least he can do is zero. But then the conclusion is also ridiculous. Birmingham and Taylor found themselves in the same paradoxical situation in which Hume found himself some 200 years earlier where his logic showed him that he could not know anything while at the same time he knew he knew a lot.

This contradiction, so concisely formulated by Birmingham and Taylor yet not recognized by them or, it seems, by their readers, should have served as a warning that something was wrong with the conceptualization underlying the thinking in this area. But it did not.

Now we can see why the Fitts lists have been impotent. To the extent that we compare, numerically, human functions to machine functions we must reach the conclusion that wherever possible the machine should do the job. This may help to explain a curious aspect in designers' behavior which

has annoyed some: an annoyance expressed trenchantly by a human factors engineer over a glass of beer thusly: "Those designers, they act as if they get a brownie point every time they eliminate a man."

Let us return to the Fitts list. They vary all over the place in length and in detail. But if we try to abstract the underlying commonalities in all of them we find that they really make one point and only one point. Men are flexible but cannot be depended upon to perform in a consistent manner whereas machines can be depended upon to perform consistently but they have no flexibility whatsoever. This can be summarized simply and seemingly tritely by saying that men are good at doing that which machines are not good at doing and machines are good at doing that which men are not good at doing. Men and machines are not comparable, they are *complementary*. Gentlemen, I suggest that complementary is probably the correct concept to use in discussing the allocation of tasks to men and to machines. Rather than compare men and machines as to which is better for getting a task done let us think about how we complement men by machines and vice versa to get a task done.

As soon as we start to think this way we find that we have to start thinking differently. The term "allocation of tasks to men and machine" becomes meaningless. Rather we are forced to think about a task that can be done by men *and* machines. The concept "task" ceases to be the smallest unit of analysis for designing man-machine systems though still remaining the basic unit in terms of which the analysis makes sense. The task now consists of actions, or better still activities, which have to be shared by men and machines. There is nothing strange about this. In industrial chemistry the molecule is the fundamental unit for many purposes and it does not disturb anybody that some of these molecules consist of hundreds, if not thousands, of atoms. The analysis of man-machine systems should therefore consist of specifications of tasks and activities necessary to accomplish the tasks. Man and machine should complement each other in

getting these activities done in order to accomplish the task.

It is possible that with a shift to emphasizing man-machine comparability new formats for system analysis and design will have to be developed, and these formats may pose a problem. I am convinced, however, that as soon as we begin thinking in proper units this problem will be solved with relative ease. Regardless whether this is so, one can now already specify several general principles that may serve as basic guidelines for complementing men and machines.

Machines serve man in two ways: as tools and as production machines. A tool extends man's ability, both sensory and motor; production machines replace man in doing a job. The principle underlying the complementarity of tools is as follows: man functions best under conditions of optimum difficulty. If the job is too easy he gets bored, if it is too hard he gets fatigued. While it is generally silly to use machines to make a job more difficult, although this may be exactly what is called for in some control situations, tools have, since their inception as eoliths, served to make a difficult job easier and an impossible job possible. Hence tools should be used to bring the perceptual and motor requirements of a task to the optimum levels for human performance. We have had a lot of experience with tools and they present few, if any, problems.

The problem is more complex with machines that do a job in place of man. Here we can return with benefit to the commonalities underlying the Fitts list. To the extent that the task environment is predictable and a priori controllable, and to the extent that activities necessary for the task are iterative and demand consistent performance, a production machine is preferable to man. To the extent, however, that the environment is not predictable, or if predictable not controllable a priori, then man, aided by the proper tools, is required. It is in coping with contingencies that man is irreplaceable by machines. This is the essential meaning of human flexibility.

Production machines pose a problem rarely posed by tools since they replace man in doing a job. They are not perfect and tend

to break down. When they break down they do not do the job. One must always then take into account the criticality of the job for the system. If the job is critical, the system should so be designed that man can serve as a manual backup to the machine. Although he will then not do it as well as the machine, he still can do it well enough to pass muster. This is another aspect of human flexibility—the ability for graceful degradation. Machines can either do the job as specified or they botch up; man degrades gracefully. This is another example of complementarity.

Planning for feasible manual backup is a difficult job in the contemporary complex systems that we are constructing. It has generally been neglected. In most simple systems explicit planning is not necessary since man's flexibility is generally adequate enough to improvise when the relatively simple machines break down. But this changes with growing system complexity.

It is here that "automation" should be mentioned. Some of you may have been bothered by the fact that automation is in the title of this paper but has, as yet, still to be introduced. The reason is rather simple. Although automation represents a significant technological breakthrough which has generated many specific problems, the allocation of tasks to men and machines being one of them, conceptually, an automated machine is just another machine, albeit radically different in its efficiency and performance characteristics. The problems that were generally latent or not too critical in the older, simpler man-machine systems became both manifest and critical, however, with its introduction. One of the most critical areas is manual backup.

We customarily design automated systems by allocating those functions which were either difficult or too expensive to mechanize to man and the rest to machines. As many articles in the literature indicate, we have looked upon man as a *link* in the system and have consequently given him only the information and means to do the job assigned to him as a link. When the system breaks down a man in a link position is as helpless as any other machine component in the

system. We have tended to design out his ability to take over as a manual backup to the system. At the same time the jobs performed by the machine have become more and more important and the necessity for a manual backup consequently greater. How to design a complex automated system to facilitate its being backed up manually is a neglected area. One thing seems certain. It will most probably call for "degradation" in design, that is, systematically introducing features which would not have been necessary were no manual backup needed. This is an important area for future human factors engineering research.

Another area of complementarity which is gaining in significance as the systems are getting more and more complex is that of responsibility. Assuming we lick the problems of reliability we can depend upon the machines to do those activities assigned to them consistently well, but we never can assign them any responsibility for getting the task done; responsibility can be assigned to men only. For every task, or for every activity entailed by the task, there must be a man who has assigned responsibility to see that the job be done as efficiently as warranted. This necessitates two things: the specification of clear-cut responsibilities for *every* man in the system and supplying the men with means which will enable them to exercise effective control over those system tasks and activities for which they are responsible. You may think that this is obvious—yes it is. But it is surprising how rare, and then how ineffective, our planning and design in this area are. Experience to date with automated systems shows that the responsibilities of the individuals involved are generally nebulous so that when something unexpected occurs people often do not know who is to do what. Even to the extent that these responsibilities are clarified with time and experience, the system hardware often makes it difficult for men to assume these responsibilities, the means for man to exercise control over the areas of his responsibility being inadequate or lacking.

The complementarity of men and machines is probably much more profound and subtle than these aspects which I have just high-

lighted. Many other aspects will undoubtedly be identified, elaborated, and ordered to the extent that we start thinking about how one complements the other. In other words, to the extent that we start *humanizing* human factors engineering. It is not surprising that the 10 years of lack of progress pointed to by Swain and Wohl (1961) were accompanied by the conceptual definition of treating man as a machine component. Man is not a machine, at least not a machine like the machines men make. And this brings me to the last point I would like to make in this paper.

When we plan to use a machine we always take the physical environment of the machine into account; that is, its power supply, its maintenance requirements, the physical setting in which it has to operate, etc. We have also taken the physical environment of man into account, to a greater or lesser extent; that is, illumination and ventilation of the working area, noise level, physical difficulties, hours of labor, coffee breaks, etc. But a fundamental difference between men and machines is that men also have a psychological environment for which an adequate physical environment is a necessary condition but is ultimately secondary in importance. This is the truth embedded in the adage: Man does not live by bread alone. The psychological environment is subsumed under one word: motivation. The problems of human motivation are at present eschewed by human factors engineering.

You can lead a horse to water but cannot make him drink. In this respect a man is very similar to a horse. Unless the human operator is motivated he will not function as a complement to machines, and the motivation to function as a complement must be embedded *within the task itself*. Unless a task represents a challenge to the human operator he will *not* use his flexibility or his judgment, he will *not* learn nor will he

assume responsibility, nor will he serve efficiently as a manual backup. By designing man-machine systems for man to do *least* we also eliminate all challenge from the job. We must clarify to ourselves what it is that makes a job a challenge to man and build in those challenges in every task, activity, and responsibility which we assign to the human operator. Otherwise man will not complement the machines but will begin to function like a machine.

And here too men differ significantly from machines. When a man is forced to function like a machine he realizes that he is being used inefficiently and he experiences it as being used stupidly. Men cannot tolerate such stupidity. Overtly or covertly men resist and rebel against it. Nothing could be more inefficient and self-defeating in the long run than the construction of man-machine systems which cause the human components in the system to rebel against the system.

Herein lies the main future challenge to human factors engineering.

REFERENCES

- BIRMINGHAM, H. P., & TAYLOR, F. V. A design philosophy for man-machine control systems. *Proc. IRE, N. Y.*, 1954, **42**, 1748-1758.
- CRAIK, K. J. W. Theory of the human operator in control systems: I. The operator as an engineering system. *Brit. J. Psychol.*, 1947, **38**, 56-61. (a)
- CRAIK, K. J. W. Theory of the human operator in control systems: II. Man as an element in a control system. *Brit. J. Psychol.*, 1947, **38**, 142-148. (b)
- EINSTEIN, A., & INFELD, L. *The evolution of physics*. New York: Simon & Schuster, 1942.
- FITTS, P. M. (Ed.), *Human engineering for an effective air navigation and traffic control system*. Washington, D. C.: National Research Council, 1951.
- SWAIN, A. D., & WOHL, J. G. *Factors affecting degree of automation in test and checkout equipment*. Stamford, Conn.: Dunlap & Associates, 1961.

(Received July 15, 1962)

A FACTOR ANALYTIC STUDY OF PERCEIVED OCCUPATIONAL SIMILARITY¹

GEORGE G. GONYEA AND CLIFFORD E. LUNNEBORG²

University of Texas

Job perceptions and "occupational stereotypes" play leading roles in many current theories of occupational choice. Case II of Andrews' A-technique was used to explore the dimensions by which occupations are perceived. Factor analysis of perceived similarity among 22 occupational stimuli yielded results corresponding directly to 5 second-order factors obtained in an earlier study which had employed different Ss, a different procedure, and different occupational stimuli. Results also shed light on college students' perceptions of popular vocational objectives and on the relationship between job perceptions and interest factors.

Job perceptions and occupational stereotypes play leading roles in most current theories of vocational choice. Some writers (Gonyea, 1961; Merwin & DiVesta, 1959) have suggested that occupational roles may be viewed largely in terms of perceived need satisfaction potential. Several recent studies (Gonyea, 1961; Grunes, 1957; O'Dowd & Beardslee, 1960; Walker, 1958) have sought to discover the dimensions by which occupations are commonly perceived. These investigations have differed from earlier studies of job perceptions (such as the long series of rankings of occupational prestige begun by Counts, 1925) in that the subject has been relatively free to choose his own dimensions, or at least to respond to several dimensions. With the exception of the study by Grunes (1957), which yielded a crude Groups \times Levels arrangement, results of these studies have generally not conformed to usual interest groupings or conventional job classifications.

For his study, Walker (1958) had college students choose from a list of 112 adjectives the 5 which they felt best described each of 10 occupations. Although there was no indication of the nature of the adjectives contained

in the list, those which were most commonly chosen seemed to describe the various occupational groups in terms of fairly distinctive personality traits.

O'Dowd and Beardslee (1960) employed the semantic differential technique to study job perceptions. They asked college students to rate each of 15 occupations on 34 bipolar rating scales. By factor analysis they identified four dimensions of job perceptions, which they called "cultured intellect," "material and social success," "cheerful sociability," and "personal and political responsibility."

Gonyea (1961) used Case III of the A-technique, the method of nonserial matching (Andrews & Ray, 1957), to explore the dimensions underlying job perceptions. For each of 30 occupational titles, the subject was asked to choose which one of the remaining 29 occupations seemed most similar. These data were used to construct a 30 \times 30 matrix of estimated intercorrelations, reflecting the extent of perceived similarities among the occupations. This was factor analyzed to produce 12 oblique factors, presumably reflecting underlying perceptual dimensions. The matrix of intercorrelations among first-order factors was also factor analyzed, yielding five orthogonal second-order factors representing more general dimensions of job perceptions. Some of the occupations were fairly uniformly perceived by all subjects, while others were apparently perceived differently by different people, presumably reflecting different needs.

The method employed in the preceding study enjoyed considerable advantage over

¹ This report is based on a paper presented at the 1961 Annual Convention of the American Psychological Association. This research was supported by grants awarded to the University of Texas Testing and Counseling Center by the Hogg Foundation for Mental Health and the University of Texas Research Institute. Appreciation is expressed to H. Paul Kelley for his valuable contributions, and to David Sohn, Paul Liberty, and Joel Levy, research assistants.

² Now at University of Washington.

the procedures of the other two studies (insofar as their purpose was to identify unknown dimensions of job perceptions) in that it left the subject entirely free to choose his own dimensions. The problems of defining dimensions in an unknown perceptual space is, of course, a problem in multidimensional psychophysics. The A-technique was chosen for the preceding investigation primarily because it appeared to be more economical (i.e., could encompass more occupational stimuli per subject hour) than other available multidimensional scaling techniques. The method of nonserial matching was selected primarily because it was the most economical of the three cases of the A-technique. However, the use of this case posed certain other problems of administration and interpretation.

On the practical side, the subjects had considerable difficulty following the instructions for nonserial matching, and data for over a third of the sample had to be discarded. A more serious problem was the frequent appearance of stereotypical doublets and triplets which were extremely difficult to interpret. It seemed to be a property of Case III of the A-technique that such a large proportion of the variance was consumed in this manner, since each subject could select only one occupation in response to each stimulus. As a result, much of the common variance was probably obscured. The second-order factors could of course be regarded as more general dimensions, but these were difficult to evaluate, since contributions of specific occupations were not clear.

In an effort to overcome these difficulties, a new study was undertaken, using Case II of the A-technique, the method of triads. In this method, the stimuli are presented in groups of three, with instructions to select from each group the stimulus that does not "belong." This procedure has the advantage of forcing the subject to match each stimulus with several others, thereby avoiding the stereotypical pairings which occurred with Case III. It also has the advantage of simplicity. However, this is obviously a much more expensive procedure, since the number of triads required for n stimuli is $n!/3!$ ($(n-3)!$ for 30 stimuli, this would be 4,060 triads). On the other hand, it is considerably

less expensive than multidimensional scaling approaches which require several presentations of each triad (e.g., Torgerson, 1952). It is also less precise.³

A second study, using a different population, different occupational stimuli, and a different method, was also indicated as a check on the generality of the findings of the first study.

METHOD

For the present study 22 occupations were selected for investigation. These included eight titles chosen to represent specific first-order factors from the original analysis (as reference tests are often used in conventional factor analyses), and four occupations which had been ambiguous (loaded on more than one first-order factor) in the original study.⁴ The remaining 10 items included occupations which are commonly listed as vocational objectives by college students and/or represent principal interest groupings derived from correlational analyses of the Strong Vocational Interest Blank (Strong, 1959).

For 22 items, there are a total of 1,540 possible triads. For the present study these were distributed systematically among 20 different forms of a Job Perception Blank, each form consisting of 77 triads of occupational titles. The triads were arranged so that every possible pair of occupations appeared approximately once in each form of the Job Perception Blank, and so that each pair appeared in every possible position and order about equally often. These 20 forms of the Job Perception Blank were randomly distributed among 2,424 University of Texas freshmen (1,491 males and 933 females), so that each triad was completed by an average of 121 subjects. The following instructions were given:

This is a test to determine how people look at jobs. Below are a number of job titles, listed in groups of three. Look at each group and decide which two of the jobs are most alike. Then cross out the remaining job, the one most *unlike* the other two. Be sure to cross out *one* job title in *each* group. Look at the following example:

SAMPLE: farmer
~~musician~~
 rancher

³ The A-technique was not designed for precise stimulus scaling, but only "to determine the number and nature of the common attributes that are used by the subjects in perceiving . . . the stimuli" (Andrews & Ray, 1957, p. 139).

⁴ In the previous analysis, using Case III of the A-technique, multiple loadings arose when different subjects used different dimensions in their job perceptions; these differences were assumed to be related to needs of the perceiving individuals (Gonyea, 1961). This interpretation is not necessarily valid with Case II.

In this group, farmer and rancher are most alike. Therefore, musician has been crossed out. Remember, you are to cross out the *one* job in *each* group which is most unlike the other two.

Responses to the Job Perception Blank were analyzed in the following manner. The first step was to construct a 22×22 *p* matrix, each cell containing the proportion of times that each pair of occupations was left together—i.e., the third member of the triad was crossed out. The *p* matrix was next converted into an estimated intercorrelation matrix by means of a simple square root transformation. This estimated correlation matrix was then factor analyzed by the principal-axes method on the University of Texas' IBM 650 Computer.

RESULTS

The factor analysis yielded seven factors, the last two of marginal significance. These factors defined the perceptual space within which the 22 occupations had been compared. It was hypothesized that this perceptual space corresponded to that described by the second-order analysis in the earlier study. Such a hypothesis implies that the perceptual spaces from both studies could be described by the same set of five factors; or, that the factors determined in this study could be transformed to a set of factors A, B, C, D, and E,

TABLE 1
HYPOTHETICAL FACTOR MATRIX H₁

Item	A	B	C	D	E
Accountant	1.0				
Artist			1.0		
Automobile Mechanic ^{a,b}		1.0			
Aviator		1.0			
Buyer ^a	1.0				
Chemist ^a					1.0
Engineer		.7			.7
Insurance Salesman	.8			.6	
Interior Decorator ^a	.7		.7		
Lawyer ^a				1.0	
Medical Laboratory Technician ^a					1.0
Office Manager ^a	1.0				
Personal Counselor ^{a,b}	.6			.8	
Personnel Manager ^a	1.0				
Physician ^{a,b}				.7	.7
Policeman				1.0	
Radio Operator ^a					1.0
Secretary	1.0				
Social Worker				1.0	
Surveyor ^{a,b}	.6	.8			
Teacher				1.0	
Writer			1.0		

^a Titles from previous study.
^b Ambiguous in previous analysis.

TABLE 2
HYPOTHETICAL FACTOR MATRIX H₂

Item	A	B	C	D	E
Accountant	1.0				
Artist			1.0		
Automobile Mechanic		1.0			
Aviator		1.0			
Buyer	1.0				
Chemist					1.0
Engineer		.7			.7
Insurance Salesman	1.0				
Interior Decorator	.7		.7		
Lawyer				1.0	
Medical Laboratory Technician					1.0
Office Manager	1.0				
Personal Counselor	.7			.7	
Personnel Manager	1.0				
Physician				.7	.7
Policeman		.6		.8	
Radio Operator		.8			.6
Secretary	1.0				
Social Worker				1.0	
Surveyor		1.0			
Teacher				1.0	
Writer			1.0		

which would closely approximate the second-order factors A', B', C', D', and E' from the previous study.

Proceeding on the foregoing hypothesis, a hypothetical factor matrix was constructed, in which the total variance for each occupation was allocated to one or more of the five factors. This hypothetical factor matrix is shown in Table 1. In general, the variance for reference items from the previous analysis (Footnote a in Table 1) was assigned in accordance with the loadings of the first-order factors in the original second-order factor matrix. For example, the variance for Chemist—which was the reference item for a first-order factor which had loaded entirely on Factor E' in the previous study—was assigned to Factor E in the hypothetical matrix. Multiple loadings indicate occupations which were ambiguous in the original first-order analysis (e.g., Personal Counselor), or which loaded on first-order factors which were ambiguous in the second-order analysis (e.g., Physician). Hypothetical loadings were assigned to new items not included in the previous study on the basis of the authors' interpretations of the dimensions obtained in the original analysis.

Using a method developed by Horst (1955),

the unrotated factor matrix obtained from the computer was next rotated so as to fit as closely as possible the hypothesized factor matrix. In general, the resulting rotated factor matrix conformed quite well to the hypothesis, though there were some minor discrepancies. To take account of these discrepancies, a new, slightly modified hypothetical factor matrix was generated, in which the variances for Insurance Salesman, Policeman, Surveyor, and Radio Operator were redistributed. This revision of the hypothetical factor matrix is presented in Table 2.

The original, unrotated factor matrix was again rotated, this time so as to approximate the revised hypothetical matrix. The results were very similar to those obtained from the first rotation, except that the factor structure was now somewhat simpler. The final, rotated factor matrix is reproduced as Table 3. The resulting factors may be regarded as the dimensions by which these 2,424 students perceived these 22 occupations. By comparing Tables 1 and 3, it may be seen how closely the final solution conformed to the original hypothesis, which in turn was based on the results of the earlier study. To a rather

remarkable extent, this analysis tended to validate the previous results.⁵

Two items from the original study—Surveyor and Radio Operator—behaved differently in the present analysis. Surveyor had been extremely ambiguous in the previous study, with loadings on two first-order factors, which in turn had loaded on three different second-order factors: A', B', and E'. The loading for E' was dropped in the initial hypothesis and that for Factor A disappeared in the final rotation. Radio Operator, on the other hand, was the reference item for a

⁵ As an independent check on the hypothesis, the original principal axes factor matrix was rotated independently to a simple structure criterion, first to an orthogonal solution by the quartimax method, then graphically to an oblique solution. When this was done, the first five factors closely resembled Table 3, and the last two emerged as residual factors, each with only one positive loading and two negative loadings of moderate size. In applying Horst's method, the variance for the two residual factors was preserved by using the original principal axes factor matrix to reproduce the correlation matrix. Three of the four occupations with negative loadings on the two residual factors (Interior Decorator, Automobile Mechanic, and Radio Operator) then appeared with secondary positive loadings in the final result.

TABLE 3
OBLIQUE FACTOR MATRIX H₂

Item	A	B	C	D	E
Accountant	.59	.00	-.05	-.11	.14
Artist	-.08	.08	.76	.07	.04
Automobile Mechanic	.12	.56	-.05	-.10	.32
Aviator	-.16	.78	.06	.04	.14
Buyer	.65	.07	.08	-.09	-.16
Chemist	-.07	.12	.04	.08	.66
Engineer	-.01	.45	.03	.01	.38
Insurance Saleman	.54	.12	-.01	.14	-.16
Interior Decorator	.39	.03	.57	-.12	-.01
Lawyer	.11	.05	-.03	.53	.04
Medical Laboratory Technician	.08	.08	-.06	.03	.74
Office Manager	.64	.00	-.12	.05	.06
Personal Counselor	.28	-.09	.06	.49	.02
Personnel Manager	.53	-.03	-.07	.23	.03
Physician	-.09	-.04	-.06	.50	.57
Policeman	-.05	.53	-.17	.59	-.10
Radio Operator	.07	.61	.04	-.09	.25
Secretary	.51	-.07	.12	-.01	.06
Social Worker	.16	.01	.11	.58	-.03
Surveyor	.06	.53	.09	.07	.09
Teacher	.07	-.08	.20	.47	.16
Writer	.10	-.06	.64	.13	-.06

first-order factor from the original analysis whose single loading on Factor E' had been difficult to interpret; in the present analysis, some of the variance for Radio Operator remained with Factor E, but most of it went to Factor B, where it seemed to better belong.

Another discrepancy between the initial hypothesis and the final solution occurred for Automobile Mechanic, which had originally loaded on Factors B' and E'. The loading for Factor E' was not retained in the hypothetical matrix because, like that for Surveyor, it was small and seemed inappropriate; however, unlike Surveyor, Automobile Mechanic continued to load on Factor E in the present study even though it was not hypothesized.

Description of Factors

Factor A included Buyer, Office Manager, Personnel Manager, and Interior Decorator, which were all reference items from Factor A'. Factor A included also a number of new items—Accountant, Insurance Salesman, and Secretary—which helped further to define this as a fairly straightforward business dimension. It seemed to include occupations from Roe's (1956) Groups II and III, Business Contact and Business Organization, suggesting that college students do not spontaneously distinguish between these two kinds of activities. Occupations in this group appear capable of satisfying such diverse needs as Order, Dominance, Conformity, and Exposition.⁶

Factor B included Aviator, Automobile Mechanic, and Surveyor, all of which had originally loaded on Factor B'. It also included Radio Operator (which in the previous study had loaded entirely on E') and two new items, Policeman and Engineer. This was evidently a technical factor, corresponding to Roe's Groups IV and V. Again her distinction between two groups was not shared by college students. These were all highly physical, masculine pursuits, perhaps reflecting manipulative or abasement needs or needs related to sex role identification.

Factor C was an esthetic factor, consisting

⁶ Since our basic theory was stated in terms of needs, we have attempted to interpret each dimension in terms of underlying motivational variables. In this we have leaned heavily on the work of Murray (1938) and Holland (1959).

of Artist, Writer, and Interior Decorator. These occupations can be characterized by such needs as self-expression, exhibitionism, and nonconformity. In the original analysis, Factor C' had contained just two doublets: Novelist-Humorist and Television Producer-Motion Picture Film Editor. This serves to illustrate the principal difficulty with the method of nonserial matching: most of the variance for these four occupations had been tied up in the two stereotypical doublets, and Interior Decorator was excluded. It was for this reason that a loading on Factor C was hypothesized for Interior Decorator, even though that occupation was originally an unambiguous reference item from Factor A'.

Factor D was a rather heterogeneous group: Policeman, Social Worker, Lawyer, Physician, Personal Counselor, and Teacher. These six occupations cut across at least three of Roe's groups and four of the Strong vocational interest areas; however, they do have in common an element of service. They are all occupations which might be expected to satisfy needs such as Nurturance, Intraception, or Voyuerism. Lawyer, Physician, and Personal Counselor had all loaded on Factor D', which had also been interpretable as a broad service dimension.

Factor E was a scientific dimension, possibly reflecting asocial interests or intellectual needs such as Cognizance or Understanding. Its highest loadings were for Medical Laboratory Technician, Chemist, and Physician, which had all previously loaded on Factor E'. Engineer, Automobile Mechanic, and Radio Operator also received moderate loadings on Factor E; Engineer was new to this study, but Automobile Mechanic and Radio Operator had also loaded on E' in the original analysis.

DISCUSSION

In general, the results of the present analysis conformed quite well to the five second-order factors obtained in the earlier study. The generality of these dimensions is attested by the differences between the stimuli, the samples, and the methods for the two studies. (Twelve titles were common to both studies. The sample for the previous study consisted of male students only from the University of

Maryland; the present study was based on both male and female students from the University of Texas.⁷ The differences in methodology have already been noted.) Where differences occurred in the present analysis, they generally helped to clarify the interpretation of the dimensions found previously.

In this connection, it may be observed that the present dimensions conformed somewhat more closely (though not exactly) to usual interest groupings and conventional job classifications than did those obtained in the previous study. In particular, the resulting factor structure resembled the occupational classification proposed by Holland (1959), although again certain occupations were misplaced and one class (Verbal Activity) failed to appear. (It is worth noting that Holland's classification was based on research involving individual responses to occupational titles.) Both Roe's groups and Strong's interest areas were largely obscured in the case of specific occupations; nevertheless, the overall pattern was interpretable along lines corresponding to conventional interest groupings.

It is interesting that both Strong (1943, pp. 135-138) and Roe (1956, pp. 147-149) have complained that many occupations were not unambiguously assignable to one or another interest area or classificatory group. When the matrix of correlations among the Strong occupational scales (Strong, 1959, pp. 36-37) is examined directly, some of the apparent discrepancies between perceptual and interest factors tend to disappear. For example, in the present study, Writer and Artist appeared together in Factor C, while Strong placed Writer (Author-Journalist) with Lawyer in Group X. Examination of Strong's intercorrelation matrix reveals that the Author-Journalist scale actually correlated more highly with Artist than with Lawyer. On the other hand, although Lawyer, Physician, and Policeman all loaded on the service dimension in the present study, none of the corresponding Strong scales correlated very

highly with each other or with any of the so-called "social service" scales.

As may be seen in Table 3, the resulting factor structure was still rather complex, indicating that some occupations were again ambiguous—i.e., loaded on more than one factor.

For example, the variance for Automobile Mechanic was divided between Factors B and E—the technical and scientific dimensions—as it also was in the previous study. Apparently some subjects perceived auto mechanics as scientists. This was a remarkably persistent misperception, which this time *reappeared* in spite of our efforts to hypothesize it away.

Radio Operator was similarly ambiguous, i.e., seen by some as technical and by others as scientific. (In the previous study, the common variance for Radio Operator had been accounted for almost entirely by the scientific dimension.)

The variance for Engineer was similarly distributed between Factors B and E. The misperception of engineering as a technical occupation, akin to auto mechanics or radio operating, is fairly common among prospective engineering students who later appear as vocational counseling clients. In accordance with the theory of job perceptions advanced previously (Gonyea, 1961), it is suggested that these individuals may have originally seen engineering as an occupation which could satisfy needs for physical activity, abasement, masculine identification, etc.; if engineering cannot meet these expectations, they then become dissatisfied.

Equally familiar to psychologists in university counseling centers is the dissatisfied premedical student who views medicine as primarily a social service occupation: 6 years of impersonal scientific training do not afford much opportunity for satisfaction of the Nurturance motive. In both this study and the previous one, Physician was predictably ambiguous, with sizable loadings on both Factors D and E, the service and scientific dimensions.

Personal Counselor—which in the previous analysis loaded ambiguously on Factors A' and D', the business and service dimensions—loaded on these same two dimensions again

⁷ Job perceptions of female subjects are being investigated separately in another study now under way, using the original 30 occupations and Case III of the A-Technique. The first-order factors for women appear very similar to those previously reported for men only (Gonyea, 1961).

in the present study. Apparently counseling is sometimes perceived as a directive activity involving dominance over other people, and at other times as a service occupation, capable of satisfying needs like Nurturance or Intraception.

Policeman was also ambiguous, with sizable loadings on Factors B and D, the technical and service dimensions. On the Strong vocational interest profile, Policeman is included in the technical interest area, while Roe and the Dictionary of Occupational Titles regard it as a service occupation; apparently college students are divided between these two views.

Finally, in this analysis Interior Decorator was regarded by some as a business occupation, capable of satisfying needs like Conformity or Exposition, and by others as an art, more related to Nonconformity and Exhibitionism.

REFERENCES

- ANDREWS, T. G., & RAY, W. S. Multidimensional psychophysics: A method for perceptual analysis. *J. Psychol.*, 1957, **44**, 133-144.
- COUNTS, G. S. Social status of occupations: A problem in vocational guidance. *Sch. Rev.*, 1925, **33**, 16-27.
- GONYEA, G. G. Dimensions of job perceptions. *J. counsel. Psychol.*, 1961, **8**, 305-312.
- GRUNES, WILLA F. Looking at occupations. *J. abnorm. soc. Psychol.*, 1957, **54**, 86-92.
- HOLLAND, J. L. A classification for occupations in terms of personality and intelligence. Paper read at American Psychological Association, New York, September 1959.
- HORST, P., & SCHAE, K. W. The multiple group method of factor analysis and rotation to a simple structure hypothesis. Technical Report, Public Health Research Grant M-743 (C1), April 1955, University of Washington Division of Counseling and Testing Services.
- MERWIN, J. C., & DiVESTA, F. J. A study of need theory and career choice. *J. counsel. Psychol.*, 1959, **6**, 302-308.
- MURRAY, H. A. *Explorations in personality*. New York: Oxford Univer. Press, 1938.
- O'DOWD., D. D., & BEARDSLEE, D. C. College student images of a selected group of professions and occupations. Report, Cooperative Research Project No. 562 (8142), 1960, Wesleyan University.
- ROE, ANNE. *The psychology of occupations*. New York: Wiley, 1956.
- STRONG, E. K., JR. *Vocational interests of men and women*. Stanford: Stanford Univer. Press, 1943.
- STRONG, E. K., JR. *Manual for Strong Vocational Interest Blanks for Men and Women*. Palo Alto, Calif.: Consulting Psychologists Press, 1959.
- TORGERSON, W. S. Multidimensional scaling: I. Theory and method. *Psychometrika*, 1952, **17**, 401-419.
- WALKER, K. F. A study of occupational stereotypes. *J. appl. Psychol.*, 1958, **42**, 122-124.

(Received June 7, 1962)

SENIORITY IN WORK GROUPS:

A RIGHT OR AN HONOR? ¹

NORMAN R. F. MAIER AND L. RICHARD HOFFMAN

University of Michigan

In the New Truck Problem, a role playing case, George, the senior man, receives the truck most often. What happens when he is absent? George received the new truck in 48.5% of 103 student groups who role played the standard case, but in 73.2% of the 41 groups where he was absent (p difference $< .01$). Bill, the next most senior man, received the truck significantly ($p < .01$) less often when George was absent and was not compensated with George's truck. Nevertheless, Bill was not more often dissatisfied. Seniority may be an honor, not a right, and not readily claimable by the senior man. Also group members are often willing to sacrifice personal gain for the group's benefit.

In management-union negotiations seniority privileges play a prominent part. Job security, periodic pay increases, and various job or shift preferences are commonly awarded on the basis of seniority (Barkin, 1954). Do rank-and-file workers accept these privileges as fair and just or are they imposed on them from past practices? How widespread is this favored treatment accepted in our culture?

Several experiments with the New Truck Problem (Maier & Hoffman, 1962; Maier & Zerfoss, 1952) have shown that when crew members have the opportunity to make the decision George, the senior worker, is given the new truck most often, even though he already has the best truck in the crew. When the case is role played with students or management personnel, in this country or in England, George stands the best chance of receiving the new truck. The next most likely to receive the new truck are Bill and John, the second and third most senior crew members. These results of role playing correspond with the writers' experiences in industry when groups of workers make decisions where seniority is involved.

Suppose that George, the senior man in the New Truck Problem, is not present in the discussion to state his case. The common sense reaction is almost unanimous that George would seldom receive the new truck.

Since all members of the crew want the new truck, people reason that George must be present to defend his rights. On the other hand, it is possible that if George did not defend his rights he would be honored for his seniority by being awarded the new truck.

These two predictions were subjected to empirical test by comparing the decisions of groups where George was present with those of groups where George was absent. It was hoped that the results of this experiment might clarify some of the social values underlying the use of seniority as a basis for awarding privileges.

METHOD

Problem. The New Truck Problem (Maier, 1955) is a six-man role playing problem involving a foreman and five telephone repairmen, each of whom drives a truck to perform his service calls. Information provided to the group shows that at the time of the meeting:

George, with 17 years in the company, drives a 2-year-old Ford truck. Bill, with 11 years in the company, drives a 5-year-old Dodge truck. John, with 10 years in the company, drives a 4-year-old Ford truck. Charlie, with 5 years in the company, drives a 3-year-old Ford truck. Hank, with 3 years in the company, drives a 5-year-old Chevrolet truck.

The foreman is told that a new Chevrolet truck has been assigned to his crew and that, since he has had difficulty in the past when he assigned new trucks to crew members, he was to allow the crew to decide who should get the new Chevrolet. Each crew member is given a short individual description of his feelings and a logical reason why he should get the new truck. In George's case he feels entitled to the truck because he has seniority and prefers

¹ The research reported here was conducted in conjunction with United States Public Health Service Grant Number M-2704 from the National Institutes of Health.

TABLE 1
ALLOCATIONS OF NEW TRUCKS TO
CREW MEMBERS

New truck given to:	Standard		George Absent	
	<i>N</i>	%	<i>N</i>	%
George (17 years)	50	48.5 ^a	30	73.2 ^a
Bill (11 years)	30	29.1 ^a	3	7.3 ^a
John (10 years)	22	21.4	8	19.5
Charlie (5 years)	1	1.0	0	0.0
Hank (3 years)	0	0.0	0	0.0
Total	103 ^b	100.0	41	100.0

^a Percentage difference between standard and George Absent conditions is significant at the .01 level of confidence by critical ratio test.
^b Three of the 106 groups in this condition failed to arrive at a decision by the end of 40 minutes.

Chevrolets. The foreman calls a meeting of the crew, presents the problem of allocating the new truck and conducts the discussion until a decision is reached.

In the variation of the case ("George Absent"), the foremen were told that George was either "home sick" or "home with a cold." They were given George's role and were told that whenever anyone in the group was ready to learn how George felt about the new truck, he was to read George's role aloud as if he had obtained this information from a telephone call to George.

Subjects. The case was multiple role played (Maier & Zerboss, 1952) in each of eight semesters of an undergraduate course in the psychology of human relations. Data for the standard version were collected from six semesters totaling 106 groups and for the George Absent version from two semesters totaling 41 groups. Differences in the distributions of solutions among semesters within each version of the problem were insignificant, justifying their combination for the present analysis.

Data Collected. Each foreman reported the person who received the new truck, the truck that was discarded, and whatever exchanges of trucks occurred

as a result of the discussion. Also, the names of workers who were dissatisfied with the decision were reported in all semesters except in one of the standard versions.

RESULTS

The number and proportion of times the new truck was given to each of the crew members under the standard conditions and in the George Absent version are shown in Table 1. Looking only at the data for the standard version of the case, we see that the new truck was given almost exclusively to the three most senior men: George, 48.5%; Bill, 29.1%; John, 21.4%. Since, in slightly more than half these cases, the new truck was given to Bill or John, factors other than seniority also played a part in the groups' decision making when George was there.

Examination of the results of the George Absent group reveals a striking shift in the allocation of the new truck. George received the new truck in 73.2% of these groups as compared with only 48.5% of the groups in the standard version, a difference significant at the .01 level of confidence. Sympathy does not seem to explain these results, since when George was described as being home sick he received the new truck in 71.4% of the groups, while when he was only home with a cold he received the truck 75% of the time. When George was not in the group for either reason he was one and one-half times more likely to get the new truck than when he was there.

George's increased receipt of the new truck came almost entirely at Bill's expense. In the George Absent condition Bill received the new

TABLE 2
WORKERS DISSATISFIED WITH DECISION

Condition	Number of groups	Number of workers	Percentage of time workers are dissatisfied ^a				
			Bill	John	Charlie	Hank	All workers
Standard	87 ^b	348	9.2	5.7	4.6	5.7	6.3 ^c
George Absent	41	164	12.2	12.2	7.3	7.3	9.8

^a Since George was absent in the George Absent condition, his "satisfaction" with the decision could not be determined. George was dissatisfied with the decision in 14.9% of the groups in the standard condition, significantly ($p < .05$) more often than any of the other crew members.
^b Dissatisfaction reports were not collected from individual workers in 16 groups in one semester.
^c 8.2% of all workers, including Georges, were dissatisfied in the standard condition.

TABLE 3
TRUCK BILL RECEIVED WHEN GEORGE
GOT NEW TRUCK

When George got new truck, Bill was given:	Standard		George Absent	
	N	%	N	%
George's truck	28	56.0	14	46.7
His own truck	16	32.0	12	40.0
Charlie's truck	2	4.0	2	6.7
John's truck	4	8.0	2	6.7
Hank's truck	0	0.0	0	0.0
Total	50	100.0	30	100.1

truck only 7.3% of the time as against 29.1% of the time under the standard condition, a difference significant at the .01 level of confidence. John, on the other hand, received the truck about equally often in the two conditions (21.4% of the time in the standard condition and 19.5% of the time in the George Absent condition).

Since Bill was the crew member most often deprived of the new truck in the George Absent condition, it is appropriate to ask whether he was more often dissatisfied with the decision in that condition than in the standard condition. Table 2 presents the percentages of times each crew member reported that he was dissatisfied with his group's decision. The proportions of dissatisfied workers, among those present for the discussion (Bill, John, Charlie, and Hank), were not very different in the two conditions. While the proportion of Bills who were dissatisfied increased slightly in the George Absent condition, the increase was well within chance expectancy.

Nor was Bill compensated for his loss by his more frequent receipt of George's 2-year-old truck. Table 3 shows that when George was given the new truck, Bill received George's 2-year-old Ford slightly *less* frequently in the George Absent condition than when George was present, although not to a significant degree. This result is especially surprising when it is realized that the number of crew members who received a different truck was greatest when George was given the new truck. Table 4 shows that in both

conditions more exchanges of trucks took place when George was given the new truck than when it was given to any other crew member.

Bill seems to be the loser when George is absent even though he is the senior man then present. In this condition (a) George gets the new truck more frequently, (b) Bill gets the new truck less frequently, and (c) Bill receives George's old truck in compensation less often. Despite these sacrifices Bill is not significantly more often dissatisfied in the George Absent condition than in the standard condition (12.2 versus 9.2%, respectively).

DISCUSSION

That seniority may be an acceptable basis for distributing benefits even among college students is confirmed by the high proportion of times the new truck was given in the present case to the man with the most seniority (George), followed in frequency by the two men with the next greatest seniority (Bill and John). The fact that seniority was given seemingly greater importance when the high seniority man was not present suggests that his presence causes the group to apply additional factors in making their decision. Observation of the role playing under standard conditions suggests that George often argues away his chance to get the new truck by insisting that it should be his by right of his higher seniority. His insistence on his rights

TABLE 4
NUMBER OF EXCHANGES OF TRUCKS AS A FUNCTION OF
WHO RECEIVED THE NEW TRUCK

New truck given to:	Standard Number of crew members getting a different truck					George Absent Number of crew members getting a different truck				
	1	2	3	4	5	1	2	3	4	5
George	0	2	22	13 ^a	13 ^a	0	2	11	9 ^a	8 ^a
Bill	0	21	3	6	0	0	2	1	0	0
John	1	12	6	2	1	0	6	1	1	0
Charlie	0	1	0	0	0	0	0	0	0	0
Hank	0	0	0	0	0	0	0	0	0	0

^a The proportion of times four or five crew members received a different truck when George got the new one is significantly greater ($p < .01$) in both conditions (52% and 56.7%, respectively) than when any of the other crew members received the new truck (17% and 9.1%, respectively).

may antagonize the other group members and cause them to give greater consideration to the claims made by Bill and John.

Seniority rights thus may be like favors earned. Whyte (1961, p. 391) has suggested that when a person grants a favor, the recipient feels a sense of obligation to return the favor at some later date. However, the granter of the favor will meet resistance if he claims the obligation directly. In the value system of workers, seniority may not be so much of a right as an honor. Too much insistence on it as a right may create resentment and cause it to decline in importance as a determiner of decisions.

George's increased receipt of the new truck when he was absent is surprising if one expects the group members to act in a selfish manner. However, as Hoffman and Maier (1959) have shown in a previous study of a problem which had immediate consequences for the group members themselves, people are quite capable of subordinating their own personal gains to obtain maximum gain for the group as a whole. In the present case, giving the new truck to George releases a relatively new truck for some other crew member, while Bill only has an old truck to be distributed if he gets the new truck. When George gets the new truck, everyone stands to gain a better truck than he is driving at present.

The subordination of personal gain to that

of the entire group also explains the lack of dissatisfaction among the Bills in the George Absent condition. When a decision has been reached with which all the other crew members are satisfied—even one in which Bill retains his old truck—he too can gain satisfaction from relinquishing his personal claims in the service of the group's benefit. Bill's feelings would then be similar to the high degree of satisfaction with the group's decision reported by the subjects in Hoffman and Maier's earlier study (1959), who contributed all the points they were "entitled to" to other more needy members of their group.

REFERENCE

- BARKIN, S. Labor unions and workers' rights in jobs. In A. Kornhauser, R. Dubin, and A. M. Ross (Eds.), *Industrial conflict*. New York: McGraw-Hill, 1954.
- HOFFMAN, L. R., & MAIER, N. R. F. The use of group decision to resolve a problem of fairness. *Personnel Psychol.*, 1959, 12, 545-559.
- MAIER, N. R. F. *Psychology in industry*. (2nd ed.) Boston: Houghton Mifflin, 1955.
- MAIER, N. R. F., & HOFFMAN, L. R. Group decision in England and the United States. *Personnel Psychol.*, 1962, 15, 75-87.
- MAIER, N. R. F., & ZERFOSS, L. F. MRP: A technique for training large groups of supervisors and its potential use in social research. *Hum. Relat.*, 1952, 5, 177-186.
- WHYTE, W. F., JR. *Men at work*. Homewood, Ill.: Irwin-Dorsey, 1961.

(Received June 14, 1962)

AUDITORY PERCEPTION OF POSITION AND SPEED¹

LESLIE BUCK

Industrial Psychology Research Unit, University College, London

2 experiments were carried out to test the ability of train drivers to use sounds for the recognition of position and speed. The procedure involved playing tape recordings of 2 steam locomotives leaving known points and traveling at known speeds to experienced train drivers in a laboratory situation. Results show that Ss were able to recognize the position recordings better than predicted by chance, and that they were able to rank the speed recordings relative to each other, but they were not able to assign accurate values to the latter in terms of miles per hour. There were no significant differences in respect of the 2 locomotives.

An assumption that is implicit in many papers on audition is that noise, and especially high intensity noise, is undesirable for the human subject. Too ready an acceptance of this point of view, however, may lead one to overlook the fact that in many tasks a great deal of information can be obtained from background noise. What to the casual observer may be heard as unwelcome disturbances, liable to distract the subject from the task in hand, may to the subject be important sources of information. Perhaps the subject perceives them subconsciously, and may not himself be fully aware of their relevance, but any attempt to "improve" the working environment by removing them results in real sensory impoverishment. Unless some substitute source of information is made available, the task will be made that much more difficult. This paper reports experiments investigating the use made of background noise in an industrial task.

In the task of driving a train it is necessary for the driver to know where he is in order to observe the railway signals at the correct times and in the correct sequence, and to stop his train at the correct points; and he must also know how fast he is traveling in order to observe timetable schedules and speed restrictions. In daylight and conditions of good visibility the driver can determine his position and speed visually, but in bad con-

ditions, and when there is no speedometer among his instruments, he becomes more dependent upon auditory and kinesthetic stimuli. These stimuli arise directly from the movement of the train, and by reflection as echoes from lineside features, and they provide a general background noise of high intensity ranging from high frequency audible vibrations to low frequency kinesthetic vibrations. Their value in the assessment of position and speed is emphasized by the drivers themselves, although in the past railway design and practice have been based almost exclusively on visual signals and information.

METHOD

The experiments were based upon the recognition of sounds recorded on the footplate of two steam locomotives, and reproduced by a tape recorder to experienced steam train drivers. The recordings were made on locomotives of different types (designated A and B) hauling passenger trains on the same main line service to and from London. For the recognition of position experiments the locomotives were recorded leaving eight stations (designated A to H in alphabetical order, not order of position along the railway) in both directions (designated up to London and down from London). The subject heard in each case the locomotive start away from the station and gather speed.

For the recognition of speed experiments a continuous recording was made on the footplate of each locomotive as it increased speed up a gradient, and the readings of the fitted speedometer were taken at intervals. In order to avoid any impression of acceleration, the prepared extracts were of only 20 seconds duration. Each extract was assigned a speed value, based upon the speedometer readings, to the nearest 5 mph.

The general qualities of sound produced by the two locomotives differ to a significant degree.

¹This paper is based on work carried out under the auspices of the Medical Research Council's Committee on Human Factors in Railway Accidents, and with the cooperation and assistance of British Railways, to whom the author is indebted.

Whereas Locomotive A produces a clear profile of sounds with clear crisp chimney beats and well defined echoes, Locomotive B produces a more homogeneous noise with less well marked features. These differences were brought out well by the recordings.

Experimental Design

Recognition of position. For the first experiment, each of nine subjects was presented with recordings of Locomotive A leaving each station taken in alphabetical order in either the up or the down direction. The number of up and down recordings was varied from subject to subject so that each of the nine possible combinations was presented once. Another nine subjects were given the same randomized sequences of ups and downs, but with station order reversed. Thus each recording was presented nine times among the 18 subjects.

For the second experiment, each of four subjects was presented with recordings of Locomotive A leaving Stations B, D, E, and G, two in the up direction and two in the down direction, followed by recordings of Locomotive B leaving the same four stations, two in the up direction and two in the down. The sequence of directions was such that among the locomotives paired by station each of the four possible permutations of directions was presented, and these were randomized among the four subjects according to a 4×4 Latin square. The same four sequences were presented with Locomotive B first and Locomotive A second, and the whole procedure was repeated for a second set of eight subjects. Sixteen subjects in all took part.

In Experiment 1 the recordings were of 2 minutes duration each, and in Experiment 2 of 1.5 minutes duration.

Recognition of speed. In the first experiment three sets of three recordings of Locomotive B were used representing speeds of 10, 25, and 40 mph, 15, 30, and 45 mph, and 20, 35, and 50 mph. Each of the 18 subjects was presented with one of the six permutations of one of the three sets.

In the second experiment one set of three recordings of Locomotive A was used, representing speeds of 20, 40, and 60 mph. Each of the six permutations of these recordings was presented to two subjects. The Sequences 20 40 60 and 60 40 20 were repeated two more times each on the remaining 4 of the 16 subjects.

Subjects and Procedure

The 34 subjects were main line drivers, all from the same locomotive shed, who agreed to take part in the experiment at the request of the investigator. They were mostly in their late fifties and early sixties, but their lengths of service as drivers on this line were much more variable, ranging from 1 year to 23. Length of service was measured from the year the driver had first signed the route knowledge card stating that he was "thoroughly acquainted" with the route and its signals. No

driver is allowed to take a train along a particular route until he has signed the relevant route knowledge card.

For the position experiment, the driver was told the locomotive and train that had been recorded and in each instance the station. He was instructed to listen to the recording to determine whether the train had been traveling in the up or the down direction. At the end of the sequence of eight recordings he was given the correct answers. Then followed the three speed recordings for which he was told the locomotive and asked to assess one at a time the speed in miles per hour. Correct answers were given at the end of the series. The subject wrote his answers one at a time on a piece of paper.

RESULTS

Recognition of Position

Table 1 shows the distributions by subject of the numbers of correct responses. For each presentation the subject could have guessed the correct or incorrect response with equal probability. For eight presentations, therefore, correct scores of 7 or 8 are better than chance ($p < .035$). For four presentations this test is not sufficiently sensitive to detect individual performance better than chance.

Taking the groups as wholes, the distributions of scores expected by chance take the form of binominal distributions with a mean of 4.0 and a standard deviation of $\sqrt{2}$ in the case of eight presentations, and a mean of 2.0 and a standard deviation of 1.0 in the case of four presentations. Standard errors were calculated from these distributions. In all cases the observed means differ significantly from the expected means. The possibility that the observed results were obtained by chance may therefore be discounted.

The results show that in Experiment 2 performance on Locomotive A was slightly better than performance on Locomotive B, but these differences are statistically non-significant using t test for related means. Seven of the 16 subjects scored equally well on both locomotives and the standard deviation of the difference is only .97.

Performance depended not only on the subject's ability, but also upon the difficulty of the particular selection of recordings with which each subject was presented. Table 2 shows the responses rearranged according to

TABLE 1
DISTRIBUTIONS OF SUBJECTS' POSITION RECOGNITION SCORES

Locomotive stations	Experiment 1		Experiment 2		
	A A to H	A B D E G	A B D E G	B B D E G	A and B B D E G
<i>N</i>	18	18	16	16	16
Score 0	—	1	—	—	—
1	—	3	—	1	—
2	—	2	7	7	—
3	—	7	8	6	1
4	6	5	1	2	3
5	3				5
6	3				6
7	5				1
8	1				—
Mean	5.56****	2.67***	2.63**	2.56*	5.19****

* $p < .015$.** $p < .008$.*** $p < .005$.**** $p < .001$.

recordings. Attempts to relate differences in stimulus difficulty to characteristics of the stimulus, including the gradient of the track at the station (inclined versus level), the locomotive (A versus B), the duration of the recording, and the number of discernible features in the recording, did not produce significant results.

The subjects' scores in the first experiment correlate positively and significantly ($p = .028$) with their lengths of service as drivers on this route, but this is not so in the second experiment. In neither case does score correlate significantly with age, although in both groups age and length of service are positively correlated.

Recognition of Speed

The subjects gave their responses to the speed recordings with much less confidence than was the case with the position recordings, and in a few instances they had to be urged to commit themselves. All but two of the responses were multiples of five, and most were overestimates (35 out of 54 in Experiment 1; 31 out of 48 in Experiment 2).

Table 3 shows that the responses to the recordings of the lower speeds tended to be less accurate and more variable than those for higher speeds. A three-factor analysis of variance (recording, presentation order, and

subject) was carried out on each of the four sets of data. The differences between the mean errors of Experiment 2 are statistically

TABLE 2
NUMBER OF RECOGNITIONS OF EACH
POSITION RECORDING

Locomotive	Experi- ment 1	Experi- ment 2	
	A	A	B
Station A up	7		
down	4		
B up	7	7 ^a	5
down	5	4	7 ^a
C up	6		
down	6		
D up	6	5	2
down	5	5	5
E up	7	4	4
down	5	5	5
F up	8 ^a		
down	9 ^a		
G up	9 ^a	8 ^a	8 ^a
down	4	4	5
H up	7		
down	5		
Number of presentations	9	8	8

^a Number of recognitions is better than chance ($p < .05$).

TABLE 3
SUBJECTS' RESPONSES TO SPEED RECORDINGS

	N	Actual speed	Estimated speed		Mean error
			Mean	SD	
Experiment 1 (Locomotive B)					
First set	6	10	24.2	15.7	+14.2
		25	30.8	15.1	+ 5.8
		40	44.2	16.7	+ 4.2
Second set	6	15	31.7	19.5	+16.7
		30	30.0	6.5	0
		45	50.8	6.1	+ 5.8
Third set	6	20	30.0	18.5	+10.0
		35	46.7	11.8	+11.7
		50	48.3	13.5	- 1.7
Experiment 2 (Locomotive A)					
	12 ^a	20	39.2	14.4	+19.2
		40	47.1	14.6	+ 7.1
		60	59.2	8.4	- 0.8

^a The responses of the four "remaining" subjects are not included here—see text.

significant ($p < .001$) and the errors are greater in respect of the recording presented third in the series ($p < .01$). The data of Experiment 1 did not yield significant results. Despite the inaccuracies, and the variability of the responses to each recording, the mean estimates vary according to the speed presented. Three-factor analysis of variance showed that the differences between the mean estimates are significant in the case of Experiment 2 ($p < .01$), and the second set of recordings of Experiment 1 ($p < .05$), but not in the first and third sets. In the latter cases, therefore, it must be concluded that the responses were no more accurate than if the subjects had made guesses. The recordings used in Experiment 2 were different from those of Experiment 1 in that the intervals between their assigned values are greater. For this reason, the estimates and errors made in response to the 10, 30, and 50 mph recordings of Experiment 1 were compared using two-factor analysis of variance. This yielded F ratios which are closer to the value of F for $p = .05$ than are those yielded by the first and third sets of data, but they are nevertheless nonsignificant. In order to compare ability to rate speeds on the two locomotives, the means of the

responses made to the 20 and 40 mph recordings of both experiments were compared using a nonrelated t test. The differences are nonsignificant. It is possible to consider the ability to rank speeds independently of the ability to rate speeds, and poor performance on the latter does not preclude good performance on the former. A subject may make a gross misjudgment of the recording presented first to him, but adjust his responses to the second and third recordings in such a way that the responses are ranked in the same order as the recordings even though they are inaccurate. If the task is regarded as one of placing the recordings in the correct sequence of low to high, the subject can be scored by computing an S value as for Kendall's tau. An S value of +3 is obtained when all three recordings are ranked correctly in respect of each other, and a value of +1 when there is only one misplacement. Table 4 shows the distributions of S values obtained in the two experiments. The distribution of S values for all the possible permutations of three has a mean of zero and a standard deviation of $\sqrt{\frac{11}{3}}$ for all values of N . In both experiments the observed

means differ significantly from the expected mean.

DISCUSSION

The train driver is able to determine his position at any time by virtue of his knowledge of the route. He has learned the sequence of the percepts with which he might be presented, and when given any one of them he is able to recognize it and its position within the sequence. In respect of visual percepts it is probably true to say that any one of them can be correctly identified from among his total repertoire, but except for a few instances, this is probably not true of auditory percepts. Sounds are not unique in the way that sights are. The task of driving does not call for performances of this level, however, since the driver knows his general whereabouts and direction of travel and recognizes each percept from among the small number with which he knows at any particular time he might be presented. It is not unreasonable therefore to use in the position experiments a task with only two possible responses, up or down. Given these restrictions the results of the experiments demonstrate that drivers are able to use auditory stimuli to determine position, and that route knowledge is not confined to one sense modality.

The ease with which any particular percept is recognized depends in part upon the adequacy of the subject's route knowledge, and this is confirmed to some extent by the finding in Experiment 1 that score is correlated to experience of the route. Recognition depends

also, however, upon characteristics of the stimulus, including, one might hypothesize, the intensity, duration, and uniqueness of the stimulus or sequence of stimuli. The failure to obtain clear statistical evidence of this was surprising. Thus the 1.5-minute recordings of Locomotive A in Experiment 2 were no less recognized than the 2-minute recordings of Experiment 1, although one might have supposed that the longer recording would have included a longer sequence of sounds having a profile more unique and more liable to be recognized. Similarly one might have expected the recordings containing the greater numbers of discernible features (as judged by a senior locomotive inspector) to have been more often recognized. The most unexpected finding, however, was that in Experiment 2 there was no difference in the performance on the two locomotives. The investigator's own observations of the locomotives at work on the railway—that they produce significantly different sound characteristics—was confirmed by drivers and locomotive inspectors, but not by the results of the experiments. This suggests that the ability of drivers to perceive sounds was more acute than first supposed.

Although statistical relationships between score and stimulus characteristics were not found, auditory inspection of the recordings indicate why some were recognized more often than others. On leaving Station G in the up direction, for example, the train enters a tunnel and produces an echo which even an untrained subject can recognize, and in fact these recordings were recognized on every

TABLE 4
DISTRIBUTIONS OF SUBJECTS' RANKING SCORES

Experiment	<i>N</i>	<i>S</i> value				Mean
		+3	+1	−1	−3	
Experiment 1						
Locomotive B	18	10	3	5	—	+1.56**
Experiment 2						
Locomotive A	16	10	4	1	1	+1.88**

** $p < .001$.

presentation. It is interesting to note that the recordings of G down which contain no such prominent feature, were not always recognized correctly by inference. Subjects reported thinking that they heard the train entering the tunnel—the recordings were very noisy even when “featureless”—which suggests that it is more difficult to hear the absence of a sound than its presence. The very small number of recognitions of Locomotive B leaving Station D in the up direction seemed to be due to the fact that the train had stopped with its locomotive ahead of the station overbridge, which therefore did not appear in the recording. By contrast Locomotive A had been stopped under the bridge and the recording clearly indicated its movement beyond it. (This meant of course that the recordings differed in respect of more than one variable, but this was unavoidable given the practical difficulties of preparing the recordings. No other similar defect was detected.) It seems that the presence of a prominent auditory feature makes a recording more easily recognized, whereas the effect of the lack of such a feature is not so clear-cut. Apart from the difficulty of measuring degree of prominence therefore, a direct correlation between prominence and recognition cannot be necessarily expected.

Some errors of recognition appeared to be due to the method used by the subject. Post-experiment interviews indicated that listening to the profile of echoes was not the only means of performing the task, but that it might also be done by listening to the way the locomotive was driven. In particular, subjects could hear from changes in the intensity of the exhaust noise the point at which the Walschaerts cutoff gear was changed, from this infer the gradient, and hence the direction of travel. There is some validity in this method since in the recordings of Experiment 1 there is a positive and significant correlation between the time elapsing before this change, and the gradient of the track. By this method recordings of stations lying on inclines should have been more easily recognized than those on the level since the differences in elapsed time between the two directions is greater. No such correlation was found.

This method of recognition is in fact misleading because of the considerable individual differences in the way the drivers drive their locomotives. The drivers of the recorded locomotives had not always done what the subject would have done in those places. This situation would not arise in real life since the driver would have to use the echo method of recognition, but it does indicate the extent to which the subjects could interpret the sounds. In Experiment 2 some attempt was made in the experimental instructions to avoid this difficulty but not, it is thought, with very much success. It was after all difficult to prevent the driver listening to the technique of his colleague on the recording, and it would have to be controlled by using a more elaborate form of response than simply “up” or “down.”

The task of estimating speed was more difficult than that of determining position. This was clear from both the manner of the subjects' responses and their analysis. Although they were able to rank the recordings relative to each other, assigning them values from an external scale was not so easy, especially in respect of the lower speeds which were grossly overestimated. This may have been due to inadequacies of the recordings—a few subjects did indicate the absence of certain features which would have been present in real life—and poor performance on recordings cannot preclude accurate performance in real life. On the other hand it may have been due to real inability on the part of the subjects to estimate speed.

Although it has been assumed that drivers can make objectively accurate speed assessments (until recently few steam locomotives were fitted with speedometers), it is questionable how accurately they make them in view of the absence of reliable knowledge of results during the period of acquisition of speed estimating skill. The fireman learning to drive has had to check his estimations against either laborious timings of the train running over known distances, or else the authoritative assertions of his driver whose own skill had been acquired under the same unsatisfactory conditions. That these estimates are not quite so accurate as sometimes suggested is indicated by the observation that

older drivers have been known to criticize as unreliable the speedometers that have been recently installed on steam locomotives.

It is conceivable, of course, that these drivers are right, and that therefore the mean errors in these experiments might have been less had the speedometers concerned been more accurate. This point does not, however, invalidate the findings of this experiment, since the statistical analyses included an analysis of the actual responses, which remain the same whatever the true value of the stimuli. No matter which of the subjects gave the most accurate responses, there were many others who were grossly inaccurate. As far as auditory estimates are concerned, therefore, speedometer readings appear to be a more reliable indication of true speed.

A further indication that drivers' speed estimating skill is not so accurate as they themselves claim, is the practice of civil engineers to set speed limits for running over

track in bad condition lower than the actual condition of the track demands. The limits are set in order to compensate for errors of underestimation rather than overestimation. In daylight, using visual stimuli, it seems that errors of speed assessment are in the opposite direction to that found in these auditory experiments (Barch, 1958).

Ratings in respect of higher speeds tended to be more accurate. This, together with the fact that the three recordings were separated by greater intervals, probably accounts for the fact that Experiment 2 produced results of greater statistical significance than those of Experiment 1. There is no evidence that the type of locomotive produced them.

REFERENCE

- BARCH, A. M. Judgments of speed on the open highway. *J. appl. Psychol.*, 1958, **42**, 362-366.

(Received July 3, 1962)

CREATIVITY TESTS AND ACHIEVEMENT IN HIGH SCHOOL SCIENCE¹

VICTOR B. CLINE, JAMES M. RICHARDS, JR., AND WALTER E. NEEDHAM

University of Utah

This study investigated the relative validities of a battery of "creativity tests" and an IQ test for predicting several indices of achievement in high school science. Criteria included grade-point average in science courses, percentile rank on the STEP Science Achievement Test, teacher rating of overall scientific potential, number of high school science courses taken, and a measure of involvement with science. Results indicated that the creativity tests did have considerable predictive validity against each criterion for each sex and that the criterion variance accounted for by the creativity tests is to a substantial degree independent of IQ. Contrary to findings of other investigators, teachers did not discriminate against highly creative pupils in their ratings.

In recent years, increasing attention has been paid by psychological researchers to an area of investigation loosely called "creativity." Since creativity has been only vaguely defined, this increased attention has many of the characteristics of a fad. However, the results of the best creativity research have suggested that many currently popular ideas about human psychology are somewhat oversimplified.

This is particularly true of current views about intellectual processes and abilities. Applied workers, both in clinical and in educational settings, often conceptualize the human intellect almost solely in terms of the IQ. Such a view of the intellect, of course, is implicitly or explicitly a general factor theory of intelligence similar to that of Spearman (1927), and as such has a considerable advantage in economy of conceptualization. In addition, it is supported somewhat by the undoubted success with which IQ tests can be used to predict school grades. This approach, however, presents difficulties when a student does either better or poorer than one would expect on the basis of his IQ. Such problem cases are typically dealt with by labeling the "offending" student an overachiever or underachiever and attributing the deviation from what would be expected on

the basis of IQ to "motivational factors." Since these motivational factors are seldom identified, it appears that such an explanation is often a picturesque substitute for a confession of ignorance.

There is an alternative approach to the intellect which has been much less influential in educational and clinical circles. This approach is to conceive of the intellect as composed of several relatively independent, or separate, capacities, or, in other words, multiple factors. In this approach, over- and underachievement are attributed to intellectual abilities not represented in the IQ. This explanation has some advantage over explanations in terms of motivation in that it is more amenable to empirical test with existing techniques.

The earliest investigator to apply this approach in a really systematic way was Thurstone (1938), who conceived of the intellect as composed of several "primary factors." More recently Guilford and his various associates (1950, 1951, 1956, 1958, 1959) have used a similar approach and extended and refined considerably the work of Thurstone. At the present time, Guilford's (1959) theory of the structure of the intellect involves 120 separate dimensions. Well over a hundred of these are *not* tapped by present day IQ tests (and nearly half have not yet been measured by any test).

Guilford feels that these dimensions may be considered functionally as well as structurally. Accordingly, there are five basic

¹ The research reported herein was performed pursuant to a contract with the United States Office of Education, Department of Health, Education, and Welfare.

operations: cognition, memory, divergent production, convergent production, and evaluation. The content which can be employed is figural, symbolic, semantic, and behavioral, while the products of the intellect are units, classes, relations, systems, transformations, and implications. Guilford further believes that creativity (or, more properly, the creative product) is a result mainly of the divergent processes, including such things as idea production, fluencies, flexibilities, and originality. These processes are typically not found in present IQ tests.

The work of Guilford, however, has concentrated most on the analysis of the relationship between tests and has been less concerned with the relationship of these tests to such external criteria as grades. As a result, critics of this approach have stated that before these tests can be said to be an improvement on conventional IQ tests it must first be demonstrated that they (i.e., creativity tests) are valid in the sense of predicting conventional criteria, and that they account for *criterion* variance that is not accounted for by IQ tests. This point is well taken, although there is some evidence for the validity of creativity tests in the work of Getzels and Jackson (1959) and of Torrance (1959).

The purpose of the present study, therefore, is to provide evidence with respect to the validity of creativity tests as predictors of performance in high school science. Specifically, this research is concerned with the answers to the following two questions:

1. Does the addition of a battery of creativity tests to a conventional IQ test result in improved prediction of performance in high school science courses?
2. Does the creativity battery alone predict performance in high school science as well as, or better than, a conventional IQ test?

METHOD

Subjects

The sample consisted of a total of 114 (74 males and 40 females) students in a Salt Lake City, Utah,² high school. Subjects were selected on the

² The authors wish to express their appreciation to Kenneth C. Farrer of Granite School District for his cooperation in providing subjects.

basis of having completed at least two courses in science beyond General Science and of having participated in an earlier study of creativity by Needham (1961). As a result of having participated in the Needham study, creativity scores were available for these students which were obtained approximately 1 year before the collection of the criterion data. Data were analyzed separately for males and females.

Predictor Measures

In the high school studied, the California Mental Maturity Inventory is routinely given to all entering students. Accordingly this test was used in this study to provide an IQ for each subject. This test has been shown to correlate highly with the Stanford-Binet.

The creativity battery consisted of the following tests developed by Guilford and was administered in the following order: Consequences, Word Association, Hidden Figures, Brick Uses, and Match Problems. The above battery is not representative of all of Guilford's creativity "elements," but rather is heavily weighted on Ideational Fluency. It should be noted that such a weighted selection of tests reduces the factorial complexity of the battery, and therefore appears to be a conservative "error" since a reduction in factorial complexity tends to reduce validity when the criterion is complex.

On the Consequences test, subjects were required to state what would happen if certain hypothetical events occurred. An example of the type of question asked is "What would happen if all the iron ore in the world disappeared?" Two scores are obtained. The first is the number of immediate consequences, a measure of Ideational Fluency. The second is the number of statements that imply that the subject was thinking in terms of more remote or indirect consequences. In answer to the above question, the response "Aluminum stock would be more valuable" would be counted in this category. This score is interpreted as measuring Originality.

On the Word Association test, subjects are required to give as many synonyms as possible for each of six words. The score is the total number of acceptable synonyms, a measure of Associational Fluency.

On the Brick Uses test, as the name implies, subjects are required to list as many different uses for a brick as they can. Two scores are obtained. The first is the total number of uses listed, a measure of Ideational Fluency. The second is the number of times a change in the *type* of use is made. An example would be a change from "To build a house" to "To throw at the umpire at a baseball game." This score is a measure of Spontaneous Flexibility.

On the Hidden Figures test, the subject is given a series of simple designs paired with complex patterns. The task for the subject is to find each simple design and its accompanying complex pattern, and the score is the total number of correct responses, a measure of Figural Redefinition.

TABLE 1
MEANS, STANDARD DEVIATIONS, INTERCORRELATIONS, AND VALIDITIES OF PREDICTORS AND CRITERIA MEANS,
STANDARD DEVIATIONS, AND INTERCORRELATIONS

Predictors	1	2	3	4	5	6	7	8	9	10	11	12	13	Female M	Female SD	Male M	Male SD
1. California Mental Maturity IQ	—	.41	.41	.26	.20	.38	.30	.35	.65	.49	.55	.52	.15	99.28	11.64	101.22	11.78
2. Word Association	.40	—	.38	.17	.04	.39	.44	.02	.29	.05	.41	.16	−.06	18.53	6.00	14.54	5.23
3. Hidden Figures	.39	.14	—	.02	.33	.38	.42	.37	.40	.31	.31	.22	−.11	19.90	4.09	19.48	5.06
4. Consequences-immediate	.16	.08	.04	—	.23	.21	.36	.26	.21	.03	.30	.22	.23	7.53	3.29	7.98	3.61
5. Consequences-remote	.37	.25	.09	.18	—	.02	.11	.17	.17	.21	.11	.15	−.12	5.15	4.89	5.13	2.72
6. Brick Uses-total	.27	.27	.21	.33	.26	—	.32	.31	.21	−.14	.19	−.01	−.02	19.83	5.57	20.37	6.47
7. Brick Uses-change	.43	.39	.25	.10	.40	.24	—	.18	.06	.03	.32	.12	.00	6.38	4.07	7.33	4.83
8. Match Problems	.40	.16	.46	.03	.20	.20	.19	—	.43	.20	.38	−.22	−.06	6.75	3.71	5.63	3.21
9. Science—grade-point average	.61	.27	.47	.02	.35	.24	.34	.47	—	.49	.76	.28	.11	2.81	.80	2.38	.79
10. STEP Science test	.47	.43	.39	.03	.43	.13	.46	.36	.48	—	.35	.35	.07	52.10	28.06	69.35	25.20
11. Teacher rating science potential	.43	.33	.36	.04	.24	.29	.31	.35	.63	.35	—	.16	.02	60.62	24.48	56.87	21.82
12. Number of science courses	.14	.23	.09	.14	.19	.15	.12	.11	.19	.38	.16	—	.42	3.37	1.36	4.74	.22
13. Involvement with science	.21	.21	.15	.22	.19	.28	.06	.27	.21	.18	.18	.38	—	1.21	.37	1.69	.65

Note.—Correlations for males ($N = 79$) are presented below the diagonal and correlations for females ($N = 40$) are above diagonal.

On the Match Problems test, subjects are given a series of drawings of groups of matches arranged to form a group of triangles. The task is to eliminate a given number of matches in such a way that a given number of triangles is left. The score is the total number of correct responses, a measure of Adaptive Flexibility.

Criteria

Five criteria relevant to achievement in high school science were used. The first of these was grade-point average in all science courses adjusted so that an average of $A = 4.00$, an average of $B = 3.00$, an average of $C = 2.00$, and an average of $D = 1.00$. Differences in the number of science courses taken were treated as error in the sense that no adjustments were made for them. The second criterion was the score obtained on the Sequential Test for Educational Progress (STEP) Science Achievement Test. This score, in the form of a percentile rank, was obtained from the permanent record of each student. The test was administered at the end of the senior year. The third criterion was a teacher rating of the "overall performance" of each student as compared to "a hundred randomly selected science students." These scores were adjusted so that 100 is the highest possible score, and the score for each student was the average of the ratings given him by all of his science teachers. The fourth criterion was the number of science courses taken up to the time the data was collected. This criterion was included because it is assumed that it reflects interest in, and involvement with, science. The fifth criterion is based on a form originally developed in connection with the research of the National Merit Scholarship Corporation.³ On this form, subjects were required to respond to the following three questions: (a) What are your major areas of interest in school? (b) What fields in the area of science do you have the greatest interest in? (c) What scientific problems do you consider to be important and of special interest to you? Do you have any notions about how these problems might be solved?

The original intention was to make detailed breakdowns of the responses obtained on this instrument. Unfortunately, however, the very poor quality of the responses precluded this. As a result, the score on this criterion is an overall rating of "involvement in science" on a 3-point scale, where 3 is the highest possible score. For each subject, the score represents the average of two independent ratings. This was done by two PhD psychologists, each rating each subject's responses.

RESULTS

The means, standard deviations, intercorrelations, and validities of the predictors

³ The authors wish to express their appreciation to John Holland for granting permission to use National Merit Scholarship Corporation forms in this research.

TABLE 2
BETA WEIGHTS AND MULTIPLE CORRELATIONS FOR MALES

Predictors	Betas without IQ					Betas with IQ				
	Science grade-point average	STEP Science test	Teacher rating science potential	Number of science courses	Involvement with science	Science grade-point average	STEP Science test	Teacher rating science potential	Number of science courses	Involvement with science
1. Word Association	.0929	.2670	.1861	.1827	.1435	.0092	.2502	.1452	.1957	.1448
2. Hidden Figures	.2987	.2629	.1965	.0608	.0545	.2099	.2457	.1542	.0761	.0581
3. Consequences-immediate	.0452	.1151	-.0121	.1706	.2197	-.0246	.1026	-.0456	.1827	.2226
4. Consequences-remote	.2206	.3125	.0662	.1922	.1904	.1266	.2935	.0208	.2075	.1934
5. Brick Uses—total	.0084	-.1848	.1268	-.0079	.1111	.0237	-.1812	.1343	-.0105	.1105
6. Brick Uses—change	.0860	.1724	.0998	-.0628	-.1700	.0327	.1779	.0751	-.0525	-.1659
7. Match Problems	.2542	.1346	.1726	.0227	.1874	.1819	.1190	.1376	.0341	.1899
8. California Mental Maturity IQ	—	—	—	—	—	.3930	.0854	.1921	-.0597	-.0092
Multiple correlation	.63	.68	.52	.32	.44	.70	.69	.55	.32	.44

Note.— $N = 79$.

and the criteria means, standard deviations, and intercorrelations for both males and females are presented in Table 1. In Table 1, correlations for males are presented below the diagonal and correlations for females are presented above the diagonal.

Multiple correlations were computed using the Fisher-Doolittle technique (Walker & Lev, 1953). These computations were made on the Burroughs 205 Datatron computer. Use of the Fisher-Doolittle technique results in a Beta weight for each predictor against each criterion in addition to multiple correlations.

Two separate analyses were made for subjects of each sex, one of the creativity battery alone and one of the creativity battery plus IQ. Results for males are presented in Table 2, and for females in Table 3.

DISCUSSION

The results presented in Table 2 and Table 3 clearly indicate that the tests in the creativity battery do have considerable predictive validity against each of the criteria for each of the sexes, and that the creativity tests do account for a substantial amount of criterion variance in addition to that accounted for by the IQ test. Furthermore, on all five criteria for males and two of the five criteria for females the multiple correlation for the creativity battery is higher than the first order validity of the IQ test, and on the remaining three criteria for females the creativity battery predicts almost as well as the IQ test. These results are consistent with the earlier findings of Getzels and Jackson (1959) that high creativity

TABLE 3
BETA WEIGHTS AND MULTIPLE CORRELATIONS FOR FEMALES

Predictors	Betas without IQ					Betas with IQ				
	Science grade-point average	STEP Science test	Teacher rating science potential	Number of science courses	Involvement with science	Science grade-point average	STEP Science test	Teacher rating science potential	Number of science courses	Involvement with science
1. Word Association	.2965	.0683	.3856	.0344	-.0848	.1485	-.0945	.2829	-.1583	-.1598
2. Hidden Figures	.2912	.3783	.0501	.3563	-.0231	.2111	.2901	-.0055	.2518	-.0637
3. Consequences-immediate	.2209	.1171	.1731	.3013	.2802	.1225	.0090	.1049	.1732	.2303
4. Consequences-remote	.0950	.1069	.0593	.1825	-.0213	.0142	.0181	.0033	.0773	-.0622
5. Brick Uses—total	-.0610	-.3386	-.1339	-.0696	.0102	-.1389	-.4132	-.1810	-.1581	-.0024
6. Brick Uses—change	-.3207	-.1293	.0435	-.0742	-.0339	-.2969	-.1032	.0560	-.0432	-.0219
7. Match Problems	.3139	.1383	.3323	-.4269	-.1161	.2233	.0327	.2657	-.5520	-.1648
8. California Mental Maturity IQ	—	—	—	—	—	.5278	.5804	.3661	.6874	.2672
Multiple correlation	.61	.47	.59	.48	.29	.74	.66	.66	.73	.36

Note.— $N = 40$.

children and high IQ children do equally well on achievement tests in spite of marked differences in IQ, and in addition help to answer the criticism of Guilford's approach that his tests have not been demonstrated to be valid. Moreover, the present study used a wide range of both creative and IQ abilities and conventional techniques for assessing validity. These results, therefore, offer persuasive evidence for the potential of Guilford's approach.

These interpretations are strengthened by the time lag between the testing and the collection of criteria data. In this regard, it should also be noted that a group IQ test of the type used in this study, because of its similarity in item forms and content to achievement tests, should therefore have a *higher* correlation with the creativity tests than would more refined IQ measures. In the opinion of the authors, therefore, this study underestimates the degree to which the creativity tests are independent of IQ.

In addition to this main contribution, this study also provides data about several important questions which were, for the purposes of this study, secondary issues. The first of these has to do with the relationship between teacher ratings and tests of IQ and creativity. Getzels and Jackson (1962) present data that suggest that teachers prefer high IQ pupils to high creativity pupils and perhaps even discriminate against high creativity pupils. Torrance (1959) also feels that high creativity pupils are less preferred by teachers. In the present study, however, there is little evidence that the teachers were penalizing high creativity pupils. The weighted combination of creativity tests predicted teacher ratings better than the IQ test alone for both sexes, and also for both sexes the addition of the creativity battery to the IQ test resulted in a substantial increase in the percentage of variance in teacher ratings accounted for.

This discrepancy between the results of this study and the results of these earlier investigators is probably due to differences in the kinds of rating the teachers were asked to make. In the present study, the teachers were rating overall performance in science, a characteristic where the "brilliant

nonconformer" is unlikely to be penalized. On the other hand, in the Getzels and Jackson study ratings included: general desirability as a student, leadership qualities, and ability to become involved in learning activities. All three of these ratings appear to have a high loading on conformity and the authors find it hardly surprising that high creativity pupils are therefore rated low on them.

The main point of all this is that the results of Getzels and Jackson have been generalized considerably without too much supporting evidence for generalizations. The results of this study suggest that Getzels and Jackson's results have been overgeneralized, and that the relationship between teacher ratings and pupil intellectual characteristics is quite complex. As with all questions in educational psychology, much more research is needed before this relationship is fully understood.

The results of this study also represent strong and interesting evidence for sex differences in science achievement. (All differences between criteria means for the two sexes were tested by means of *t* tests and all with the exception of the teacher rating criterion were significant at beyond the .01 level.) It is hardly surprising that males take more science courses than females and get more involved with science. However, it is more puzzling to find that girls get higher grades in science and tend slightly to be rated higher by teachers on scientific potential, but obtain significantly *lower* scores on an objective test of knowledge of science. An explanation in terms of teachers favoring conformity immediately suggests itself, but such an explanation cannot be advanced with any confidence in view of the relationships between the creativity tests and teacher ratings. There is, however, a clear indication that for unknown reasons *males* are penalized by science teachers (mostly male) both in grading and in rating whereas females are rewarded.

There is also some suggestion of potentially important sex differences in the relationship between intellectual characteristics and the various indices of achievement in science. This is most striking in the case of the number of science courses taken. There seems

to be a strong tendency for high IQ females to take science courses, but much less tendency for high creativity girls to take science courses. Such a difference does not appear to exist for males, and could even be reversed. An explanation again is readily suggested in terms of middle-class sex roles and the channels of expression of creativity acceptable for the two sexes, although again any specific evidence that *this* explanation is correct is lacking.

REFERENCES

- GETZELS, J. W., & JACKSON, P. W. The highly intelligent and the highly creative adolescent: A summary of some research findings. In C. W. Taylor (Ed.), *1959 Conference on the Identification of Creative Scientific Talent*. Salt Lake City: Univer. Utah Press, 1959.
- GETZELS, J. W., & JACKSON, P. W. *Creativity and intelligence*. New York: Wiley, 1962.
- GUILFORD, J. P. Creativity. *Amer. Psychologist*, 1950, 5, 444-454.
- GUILFORD, J. P. The relation of intellectual factors to creative thinking in science. In C. W. Taylor (Ed.), *1955 Conference on the Identification of Creative Scientific Talent*. Salt Lake City: Univer. of Utah Press, 1956.
- GUILFORD, J. P. Basic traits in intellectual performance. In C. W. Taylor (Ed.), *1957 Conference on the Identification of Creative Scientific Talent*. Salt Lake City: Univer. Utah Press, 1958.
- GUILFORD, J. P. Intellectual resources and their value as seen by scientists. In C. W. Taylor (Ed.), *1959 Conference on the Identification of Creative Scientific Talent*. Salt Lake City: Univer. Utah Press, 1959.
- GUILFORD, J. P., WILSON, R. C., CHRISTENSEN, P. R., & LEWIS, D. S. A factor analytic study of creative thinking: I. Hypotheses and description of tests. Report, 1951, University of Southern California, Psychological Laboratory.
- NEEDHAM, W. E. *Biographical predictors of creativity, intelligence, and teacher preference in a high school*. Unpublished master's thesis, University of Utah, 1961.
- SPEARMAN, C. *The abilities of man*. New York: Macmillan, 1927.
- THURSTONE, L. L. Primary mental abilities. *Psychometr. Monogr.*, 1938, No. 1.
- TORRANCE, E. P. Explorations in creative thinking in the early school years: A progress report. In C. W. Taylor (Ed.), *1959 Conference on the Identification of Creative Scientific Talent*. Salt Lake City: Univer. Utah Press, 1959.
- WALKER, HELEN, & LEV, J. *Statistical inference*. New York: Holt, 1953.

(Received July 9, 1962)

EFFECTS OF MAGNIFICATION ON A SUBMINIATURE ASSEMBLY OPERATION¹

RAVI M. NAYYAR AND J. RICHARD SIMON

University of Iowa

This experiment was designed to investigate the effects of magnification on the duration of elements of a subminiature assembly operation. The task consisted of grasping a metal dot .010 in. in diameter with a tweezers, transporting it to a hole, and dropping it into the hole. Ss used a binocular type industrial microscope and performed under 3 magnifications, 20X, 30X, and 40X. Precision of the task was varied by changing the diameter of the hole into which the dot was assembled. Results indicated that no single magnification was optimum for all elements. For pick up and travel loaded, 30X was optimum while for travel unloaded, 20X was optimum. Results for the assemble element were inconclusive. There was no evidence that the optimum magnification is dependent upon the precision requirements of the task.

One of the striking characteristics of space age technology has been the trend toward miniaturization of electronic and other functional systems. The demand for miniaturized parts has confronted industry with many new problems. Although fabrication of these parts can be done by automatic precision tools, the majority of assembly work must be performed manually. Since the parts are so small and the manipulations required are so precise, assembly operators must utilize magnifying devices such as the industrial microscope in order to achieve the necessary visual control of their movements.

The purpose of this investigation was to study the effects of magnification on the duration of the elements of a subminiature assembly operation. The following questions were of interest: Is there an optimum magnification for subminiature assembly work? Is the same magnification optimum for all elements of the task? Does the optimum magnification depend upon the precision of the task? Can one predict from the characteristics of the task what the optimum magnification may be? A literature survey failed to disclose a single systematic study concerned with the effects of magnification on subminiature skills.

¹ This article is based on a master's thesis done by Nayyar under the direction of Simon. The authors are indebted to Dee W. Norton for his advice on the statistical analysis.

METHOD

To study the effects of magnification on the elements of a subminiature assembly operation, a task was designed which, in terms of operations performed and demands upon the operator, was similar to actual operations involved in industry. The task consisted of grasping with a tweezers a spherical metal dot .010 inches in diameter (roughly half the size of a grain of sugar), transporting the dot about .20 inches to a hole, and dropping it into the hole. Subjects viewed the task through a binocular microscope and performed the entire task under three magnifications: 20X, 30X, and 40X. The precision of the task was also manipulated by varying the diameter of the hole into which the dot was assembled. The duration of the elements of the task under each experimental condition was automatically recorded on four electronic timers.

Apparatus

Figure 1 shows the specially designed assembly jig used in this study. The jig was a cylindrical piece of aluminum $\frac{3}{8}$ inch in diameter and $\frac{1}{4}$ inch high. The top face of the jig was made with a lip around its circumference so as to hold a supply of about 40 dots within its surface area. The cylinder was cut in half along its axis, and the two halves separated by a nonconducting material, so that each half of the jig would activate a different timer relay. The separator projected above the surface of the face just enough to keep the dots from rolling from one half of the jig to the other.

The right half of the jig face was used as a receptacle for the dots. The left half contained the hole into which the dots were dropped. Four hole sizes were used in this experiment. They were .020 inch, .035 inch, .051 inch, and .063 inch in diameter or about 2, $3\frac{1}{2}$, 5, and $6\frac{1}{2}$ times the

size of the dot. Four brass inserts ($\frac{1}{16}$ inch diameter) were made; each contained a different sized hole. Task precision was altered by simply placing a different brass insert into the hole in the left half of the jig. The inserts fitted snugly into the hole so as to maintain electrical contact.

A wooden fixture kept the jig centrally positioned within the magnified field of an American Optical Company binocular type industrial microscope. This instrument was equipped with interchangeable ocular and objective lenses to obtain different magnifications. Magnification was varied by using three different objective lenses: 2X, 3X, and 4X. The combination of each different objective with a 10X ocular lens produced the 20X, 30X, and 40X magnifications used in this study. A spotlight provided 475 foot candles of uniform illumination within the magnified field.

The height of the work table and the chair, and the positioning of the back rest of the chair were fixed in accordance with recommended dimensions (Barnes, 1958). A 37-inch-high table with a comfortable foot rest was used. The chair height was adjusted to allow 9.5 inches of space between the top of the chair seat and the under surface of the table.

A four-channel universal motion analyzer (Smader & Smith, 1953) and Hunter KlocKounters were used to time separately and automatically the duration of each of the four elements of the assembly operation. The four elements of the task were: pick up, travel loaded, assemble, and travel unloaded. The subject's movements in performing the task automatically activated the KlocKounters. When the subject touched the right half of the jig with the tweezers to pick up a dot, the first KlocKounter began to run and continued timing as long as the contact was maintained. When the subject lifted the tweezers and began to move the dot toward the hole, the first KlocKounter stopped and the second KlocKounter began to record the travel loaded time. When the subject touched the left half of the jig with the tweezers to assemble the dot, the second KlocKounter stopped, and the third KlocKounter began to record the assemble time. When the subject broke contact with the left half of the jig and began to move back to pick up another dot, the third KlocKounter stopped and the fourth KlocKounter began to record the travel unloaded time.

The task consisted of five cycles, i.e., picking up, transporting, and assembling five dots. At the end of the fifth assembly the subject was instructed to touch the stop plate (about .5 inch from the assembly jig) with a free outward motion of the tweezers. Contact with the stop plate stopped all four timers. Consequently, the task consisted of five pick up elements, five travel loaded moves, five assemble elements, four travel unloaded moves, and one move to touch the stop plate. The duration of the last move from the assembly area to the stop plate was recorded on the travel unloaded KlocKounter.

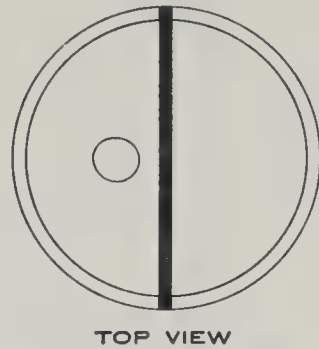
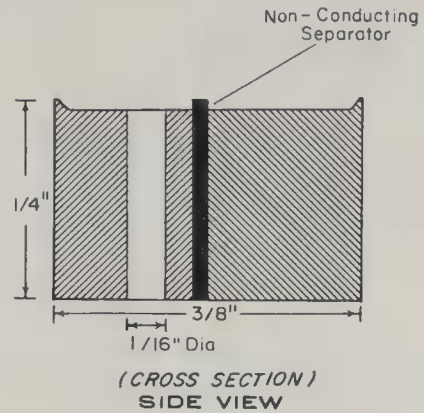


FIG. 1. Assembly jig.

Subjects

Twenty-four female students from the elementary psychology class at the State University of Iowa were used as subjects. The restrictions imposed were that the subjects have normal vision, with or without glasses, and that they be right handed. Female subjects were used because, generally, females are employed in industry on subminiature assembly work.

Experimental Design

Six subjects were assigned at random to each of the four precision conditions. Each subject performed the task under all three magnifications. The effect of order of presentation of the three magnification conditions was counterbalanced by using a 3×3 Latin square which was replicated twice in each precision condition. In other words, within each precision, two of the six subjects were assigned to each of the three orders of magnification. Two experimenters were used, each handling one of the replications. The design, basically a Type IV (Lindquist, 1953), makes possible the control of individual differences not only in evaluating the main effects of magnification and order but also in evaluating the interactions of these variables with precision.

Procedure

The subjects served in three experimental sessions separated by at least 24 hours. In each session they

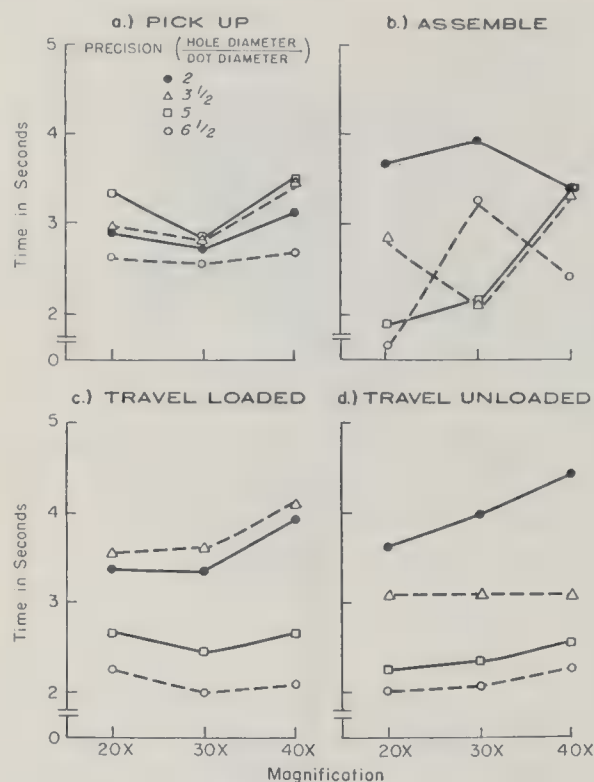


FIG. 2. Effect of magnification on duration of elements.

performed in the order and precision conditions to which they were initially assigned. The first two sessions were essentially training sessions,² and only data from the third session were subjected to statistical analysis. By the third session, performance was relatively stable, though some practice effect was still present in the assemble and pick up elements.

In the first session, subjects were given a brief introduction to the purpose of the investigation, after which the method of focusing the microscope and the manner of holding the tweezers were explained and demonstrated. This was followed by 10 familiarization trials which consisted of alternately touching the tweezers to the left and then the right half of the assembly jig, repeating this cycle five times, and then touching the stop plate. In addition to providing practice in moving within a magnified field, these trials oriented the subjects to the location of the stop plate which was outside the magnified area. The method of performing the assembly task was then explained and the subject was given five trials under each magnification. During each of the final two sessions, the subject was given 10 trials under each magnification.

RESULTS

Performance of 24 subjects during the third session was analyzed to evaluate the

² A pilot study using six subjects indicated that, after two practice sessions, performance data were reliable enough to be used for statistical analysis.

effects of magnification and precision on the duration of the elements of the task. For each subject, a median time was determined for each element under each experimental condition. The results of the analyses are described below.

Pick Up

The effect of magnification on the pickup element is shown in Figure 2a. It can be seen that the mean pick up time was least under 30X for all precision conditions. An analysis of variance³ indicated that the effect of magnification on pick up time was significant at the 1% level. Subsequent *t* tests indicated that performance under 30X was significantly ($p < .05$) faster than under 20X or 40X. These latter two conditions did not differ from each other.

The effect of precision on pick up time was not significant. In other words, increasing the general difficulty of the task had no effect on the pick up element. The effect of order was significant at the 5% level. Inspection of the data revealed a steady improvement in performance with successive trials; i.e., the significant order effect indicates the presence of a practice effect at the time of the observations.

Assemble

An analysis of variance of the assemble times revealed a significant interaction between order, magnification, and precision. In other words, the assemble time under a magnification and precision treatment combination was not independent of the order in which the magnifications were administered. This meant that no meaningful analysis could be carried out until all effects arising from the fact that each subject was given a number of magnifications in succession were eliminated. To accomplish this only those data obtained from the first magnification treat-

³ A summary of this analysis along with four additional analyses of variance has been deposited with the American Documentation Institute. Order Document No. 7501 from ADI Auxiliary Publications Project, Photoduplication Service, Library of Congress, Washington 25, D. C. Remit in advance \$1.25 for microfilm or \$1.25 for photocopies and make checks payable to: Chief, Photoduplication Service, Library of Congress.

ment administered to each subject were retained. For example, if a subject had performed the magnification conditions in the order 20X, 30X, and 40X, the data obtained from 20X were retained and the data from 30X and 40X were discarded. This procedure reduced the available data by two thirds, leaving observations for two independent subjects for each treatment combination.

These remaining data were analyzed as a two-factor simple-randomized experiment. Because of the reduced power of the test, the effect of magnification on assemble time was not statistically significant. However, there was a tendency for assemble time to increase with an increase in magnification (see Figure 2b) although the results were not nearly as systematic as for the other elements of the task. This rather confused picture is probably due in part at least to the small number of observations per cell.

Despite the reduced number of observations the effect of precision on assemble time was significant at the 10% level; i.e., an increase in precision or task difficulty tended to produce an increase in the assemble time.

Travel Loaded

The effect of magnification on the travel loaded element is shown in Figure 2c. The mean time for this element tended to be least under a magnification of 30X. An analysis of variance indicated that the effect of magnification was significant at the 10% level. Subsequent *t* tests indicated that performance at 40X was significantly (*p* < .05) slower than at 30X. There was no difference in performance between 20X and 30X, or between 20X and 40X.

The effect of precision on travel loaded time was significant at the 5% level. Except in the case of the 3½ precision condition, where two subjects gave very high time values, a decrease in the precision of the task produced a reduction in the time taken for the travel loaded move.

Travel Unloaded

The effect of magnification on the travel unloaded element is shown in Figure 2d. There was a systematic tendency for the travel unloaded time to increase with an increase in magnification. An analysis of variance indicated that the effect of magnification was significant at the 5% level. Subsequent *t* tests indicated that the performance at 40X was significantly (*p* < .05) slower than at 20X. There was no difference in performance between 20X and 30X, or between 30X and 40X.

The effect of precision on travel unloaded time was significant at the 1% level; i.e., decreasing the precision of the task produced a systematic reduction in the time taken for the travel unloaded move. In this analysis the effect of experimenter was also found to be significant at the 5% level.

Summary of Effects of Magnification and Precision

Table 1 summarizes the effect of magnification on the duration of the four elements of the assembly cycle. The table shows the mean time for each element under each magnification condition (averaged over all precision conditions). It also shows the percent slowing from the optimum magnification for each element. Disregarding for the moment the

TABLE 1
SUMMARY OF EFFECTS OF MAGNIFICATION ON DURATION (in seconds) OF ELEMENTS

Magnifi- cation	Pick up		Assemble		Travel loaded		Travel unloaded	
	Time	Slowing (%)	Time	Slowing (%)	Time	Slowing (%)	Time	Slowing (%)
20X	2.90	9.98	2.51	—	2.95	3.72	2.74	—
30X	2.64	—	2.79	11.20	2.84	—	2.88	5.17
40X	3.10	17.52	3.13	24.69	3.17	11.60	3.09	12.58

Note.—All time values are for five cycles.

TABLE 2
SUMMARY OF EFFECTS OF PRECISION ON DURATION (in seconds) OF ELEMENTS

Precision ^a	Pick up		Assemble		Travel loaded		Travel unloaded	
	Time	Slowing (%)	Time	Slowing (%)	Time	Slowing (%)	Time	Slowing (%)
2	2.84	12.03	3.63	48.74	3.54	66.60	4.01	87.91
3½	3.01	18.26	2.72	11.47	3.73	75.67	3.08	44.52
5	3.11	22.05	2.45	—	2.57	21.06	2.38	11.26
6½	2.55	—	2.44	—	2.12	—	2.14	—

Note.—All time values are for five cycles.

^a Precision is specified in terms of ratio of hole diameter to dot diameter.

statistical significance of the observed differences, it will be noted that performance was uniformly slower for 40X on all elements of the task. For pick up and travel loaded, 30X was the optimum power; for assembly and unloaded travel, 20X was optimum.

Table 2 summarizes the effect of precision on the duration of the four elements of the assembly cycle. The table shows the mean time for each element under each precision condition (averaged over all magnifications). It also shows the percent slowing from the fastest time for each element. It will be noted that an increase in the precision requirement of the task produced a fairly systematic slowing in all elements with the exception of pick up where the differences were not significant.

Subjects' Preferences

At the end of the experiment, the subjects were asked to rate the different magnifications in order of preference. Sixteen subjects preferred 20X; the remaining eight preferred 30X. All subjects rated 40X as least preferred.

DISCUSSION

This study clearly showed that duration of the elements of a subminiature assembly operation depends upon the magnification used to view the task. The same magnification was not optimum for all elements. For the pick up and travel loaded elements, a magnification of 30X was optimum; for the assemble and travel unloaded elements, 20X was optimum. For all elements performance under 40X was poorest.

Since the effects of magnification are different for different elements, the choice of a magnification to use in an industrial situation will involve some compromise unless, of course, only one element is to be performed under magnification. The optimum magnification for a task will depend upon the absolute time values involved for each element. Selection of a magnification should be made so that a minimum total task time is obtained.

Increasing the precision of the task by reducing the hole size into which the dot was assembled produced a systematic slowing of the travel loaded and travel unloaded elements as well as the assemble element. In other words, changing the precision of the assemble element altered the duration of travel movements which preceded as well as followed it. Similar results have been demonstrated by Simon and Simon (1959) in a dial setting task and by Schmidtke and Stier (1961) in a positioning task. There was no effect of precision on the pick up element, perhaps because pick up was separated from the assemble element by the intervening travel movements.

There was no evidence from this study that the optimum magnification is dependent upon the precision requirements of the task. In other words, in no element was there a significant Magnification \times Precision interaction. It is unfortunate that the interaction of order, magnification, and precision occurred in the assemble element (thus reducing the available data and the power of the test) since it was in this element

that the Precision \times Magnification interaction might most logically have occurred.

There are several aspects of this study which warrant further investigation. First, it would be of interest to investigate the possible interaction of magnification and task size. In the present study, for example, 30X was the optimum magnification for picking up a .010 inch dot. Would a higher power (perhaps 40X) be optimum for picking up a smaller dot and would a lower power (perhaps 20X) be optimum for picking up a larger dot?

Second, results of this study indicate that travel movements are slower when performed under high power. This finding should be confirmed by further research. If, in fact, movements are slower through a highly magnified field, it would be of some theoretical interest to determine the reason for this effect. The slowing of travel movements may be a function of the magnification variable per se, or it may be due to a reduction in the size of the field. As a means of investigating this question, the possibility of vary-

ing magnification without altering the size of the field is now being explored. As a practical matter, however, which of these two variables accounts for the effect is of little concern since a reduction of field size always accompanies the use of a higher magnification.

REFERENCES

- BARNES, R. M. *Motion and time study*. New York: Wiley, 1958.
- LINDQUIST, E. F. *Design and analysis of experiments in psychology and education*. Boston: Houghton Mifflin, 1953.
- SCHMIDTKE, H., & STIER, F. An experimental evaluation of the validity of predetermined elemental time systems. *J. industr. Engng.*, 1961, 12, 182-204.
- SIMON, J. R., & SIMON, BETTY P. Duration of movements in a dial setting task as a function of the precision of manipulation. *J. appl. Psychol.*, 1959, 43, 389-394.
- SMADER, R. C., & SMITH, K. U. Dimensional analysis of motion: VI. The component movements of assembly motions. *J. appl. Psychol.*, 1953, 37, 308-314.

(Received July 9, 1962)

EFFECT OF REFERENCE MARKS ON THE DETECTION OF SIGNALS ON A CLOCK FACE¹

JANE F. MACKWORTH

Defence Research Medical Laboratories, Toronto, Canada

42 female Ss were employed in each of 3 studies undertaken to determine the effect of white marks on a black clock face on the detection of signals, consisting of brief pauses of the clock hand. The signals were presented at intervals ranging from 5-14 sec.; 0-30 marks were used. The addition of 1 mark reduced the percentage of missed signals to half that of the blank face ($p < .01$). Least signals were missed when they were near the white mark. Conclusions are the detection of a brief pause in a clock hand is improved by the addition of reference marks and there is a rapid decrease in detection of frequent signals as the run continues.

Mackworth (1963) described some vigilance experiments using a clock test, in which the signal was a brief pause in the movement of the second hand of an electric clock. The clock was divided into 10 black and 10 white segments, and it was found that twice as many signals were missed when they were expected anywhere on the clock face as when they were expected only on the black segments. This was true even when the signals were given in rapid succession in a 6-minute pretest. Further, a marked decrement with time was found in the percentage of signals detected, even within the course of the pretest.

In view of the marked effect of eliminating signals from the white segments, a study (Experiment I below) was made of the effect of varying the number of white marks on the clock face, while the signals were presented only on the black segments. Experiment II was designed to see if the effect of the marks was due solely to the elimination of signals from part of the clock face; comparisons were made among various numbers of marks and also among different distributions of signals. It became clear from these experiments that two different factors were involved because even when the signals occurred anywhere on the clock face the presence of marks greatly increased the detection

of signals. Experiment III was therefore designed to investigate whether the position of a mark influenced the detection of signals.

In all these experiments the results were analyzed to reveal temporal changes in signal detection that might occur even though the signal presentation rate was high.

EXPERIMENT I

Method

The subjects sat alone in dimly lighted booths and watched a televised picture of a clock. The clock face was black except for the white marks described below, and it carried only the second hand, which was also black with a white tip. White auditory noise of 70 decibels was present throughout the trials. Seven subjects were tested simultaneously, all receiving the same signal.

Six different clock faces carrying white marks or segments ranging in number from 0 to 30 (as shown in Figure 1a) were used. The white marks were distributed regularly on a ring around the clock face. A gap of 4 millimeters separated the white tip of the hand from the segmented ring. The marks represented 3 seconds duration except for one of the two faces with 4 marks and of course the face carrying 30 marks. In these two cases the marks represented a duration of 1 second each. Thus the face with 30 segments of 1 second each (30×1) and the 10×3 face both showed white on half the circumference. The subjects were told that signals would occur only on the black segments of the clock face (Black instructions).

There were six groups of seven subjects each, and each group received all six clock faces in one session, in an order chosen from a Latin square, so that each clock face appeared once in each successive trial for a different group of subjects. The 42 subjects were housewives.

The subjects received 40 signals on each clock face in about 6 minutes; between the presentation

¹ Defence Research Medical Laboratories Project No. 234, Report No. 234-15, PCC No. D77-94-20-42, H. R. No. 237.

Grateful thanks are due to G. E. Boyes and E. A. Singer who carried out the testing.

of each face there was an intermission of about a minute. The signals were pauses of 350, 360, 370, or 380 milliseconds each; a signal sequence included 10 signals of each duration, presented randomly. The total session lasted about 45 minutes.

Results

Figure 1a shows that the percentage of signals missed from each display was related to the ratio of white to black segments; i.e., the proportion of time when no signals were expected. The results were analyzed in terms of the display, the order of the trials, and

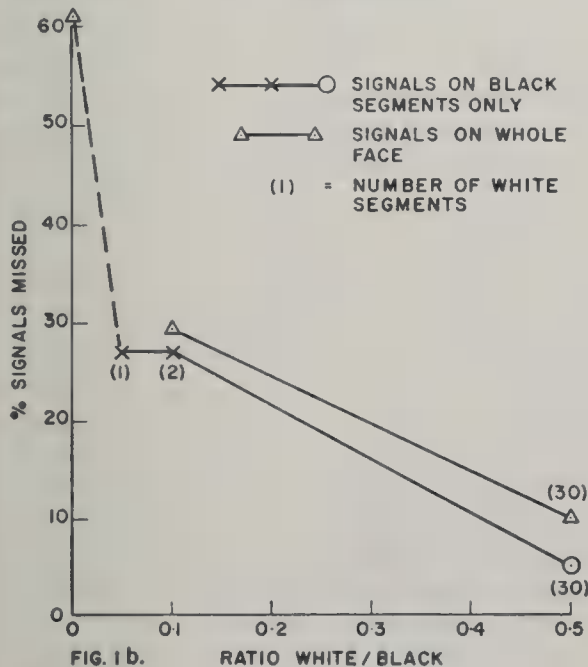
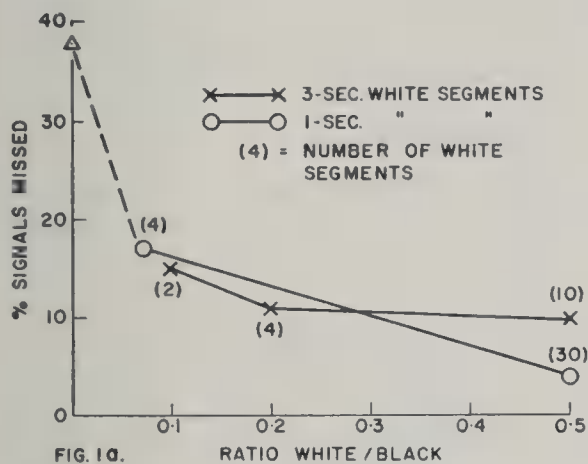


FIG. 1. The percentage of signals detected in relation to the ratio of white segments to black segments on the clock face. (Figure 1a shows signals expected on the black segments only in Experiment I. Figure 1b shows signals expected either on the black segments only or on the whole face in Experiment II.)

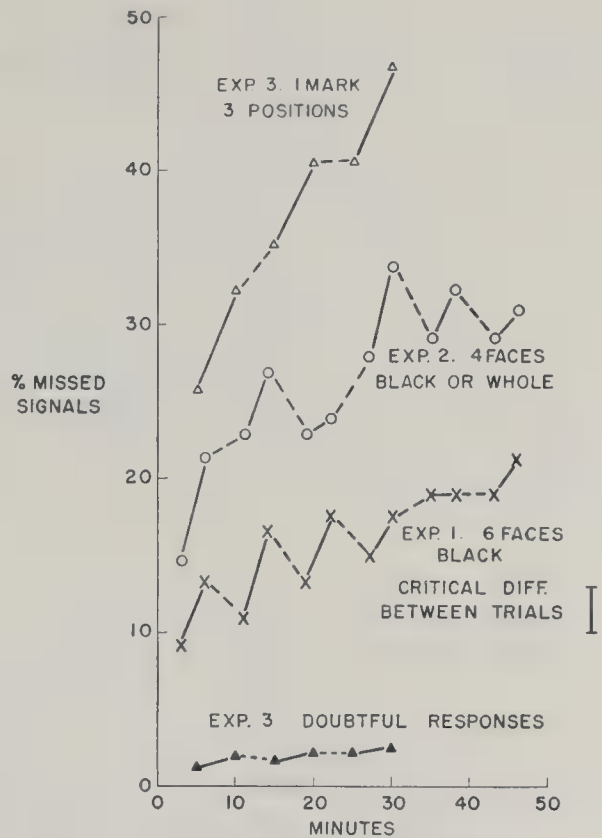


FIG. 2. Deterioration of detection during session. (The halves of each trial are connected by solid lines, and the pauses while the clocks were changed are indicated by dotted lines. Black indicates that signals occurred only on the black parts of the clock face while Whole indicates that the signals occurred anywhere on the clock face.)

the halves of each trial. Analysis of variance showed that all these main effects were significant ($p < .01$). There were significant differences ($p < .01$) among all displays with one exception: between 4 and 10 white segments of 3 seconds each. More signals were missed with the 4×1 display than with the 4×3 , suggesting that the amount of white (i.e., the proportion of time during which no signal was expected) was important when the number of marks was kept constant. On the other hand more signals were detected from the 30×1 display than from the 10×3 , showing that when the amount of white was kept constant, the number of marks was important.

The data, divided into halves of successive trials, are shown in Figure 2; each point represents 20 signals on each of 6 displays for 42 subjects. Each trial shows a greater percentage of missed signals than the previous

one. A decrement is noted also within each trial, with some recovery in the intervals between trials.

EXPERIMENT II

The experiment reported previously (Mackworth, 1963) showed that with 10 white marks of 3 seconds each, the number of signals missed when signals were expected anywhere on the clock face was double the number missed when they were expected only on the black segments. Experiment II was designed to distinguish between the effect of the number of marks on clock face and that of the expected location of signals.

Method

The two main variables in this experiment were the number of white marks and the expected location of signals. For three of the six conditions, subjects were told that signals would occur only on black segments (Black instructions) while for the other three they were told that signals would occur anywhere on the clock face (Whole instructions). The four clock faces used are indicated in Figure 1b. Two of the faces, the 30×1 and the 2×3 , were used with both sets of instructions. The 1×3 was used only with the Black instructions while the blank face was used only with the Whole instructions.

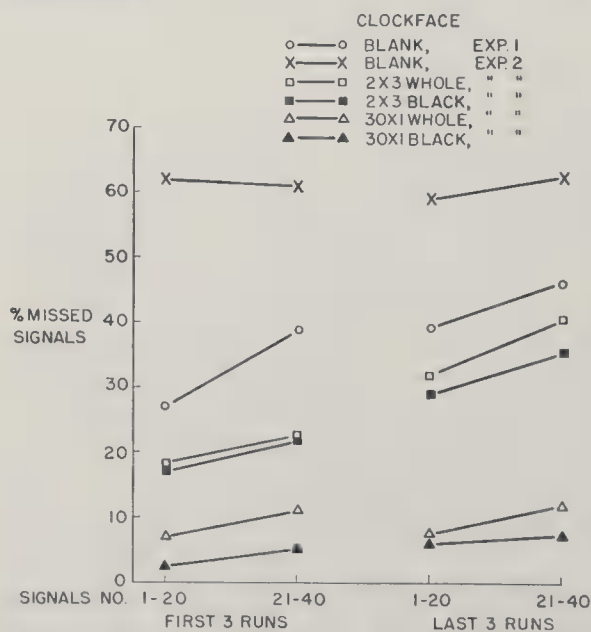


FIG. 3. Relation of temporal deterioration to display—Experiment II, and blank face, Experiment I. (Each run lasted about 6 minutes. The figure shows the pooled data for the first halves of the first three runs and similarly for the second halves of the first three runs, and the first and second halves of the last three runs.)

Six groups of seven subjects each were employed. Three of these groups received first the Whole instructions with the 30×1 , the 2×3 , and the blank clock faces; and then the Black instructions with the 30×1 , the 2×3 , and the 1×3 clock faces; while the other three groups received the instructions in the opposite order. Within each instruction order each group received the three clocks in a different order. The subjects were instructed at the beginning of each set and at the beginning of each display where they should expect the signals.

It was found in Experiment I that results were similar for all four signal durations; therefore, for simplicity, only the shortest duration of 350 milliseconds was used in this experiment. The interval between signals varied by a random design from 5 seconds to 14 seconds. There were 40 signals in a trial and each trial took approximately 6 minutes. The total session lasted about 45 minutes.

Results

Figure 1b shows that the percentage of signals missed from each display condition was again related to the ratio of white to black, even when signals were expected anywhere on the clock face. Analysis of variance was carried out in terms of the displays and the successive trials. The main effects of trials and displays were significant ($p < .01$). The percentage of missed signals was greatest for the Blank display, and greater for the 2×3 displays than for the 30×1 displays. The interaction between the displays and the trials was also significant ($p < .01$) however, and separate analyses were therefore carried out on the Black and Whole conditions for each display. The difference between trials was significant ($p < .01$) for the 2×3 display, but not for the 30×1 display; on the other hand, the difference between instruction conditions was significant ($p < .05$) for the 30×1 display but not for the 2×3 display. Figure 3 shows these differences. By comparing the first three trials with the last three, it can be seen that there is a much greater decrement for the 2×3 displays than for the 30×1 displays. On the other hand, performance on the 30×1 Black is better than that on the 30×1 Whole regardless of the trials, but this is not true for the 2×3 displays.

Figure 2 shows the successive trials and halves of each trial for all displays and conditions. Here again there is a decrement within each trial with a tendency to recovery

TABLE 1
ANALYSIS OF VARIANCE: EXPERIMENT III

Source	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>
Between subjects		3628.7		
Groups	2	101.3	50.6	
Error (b)	39	3027.4	77.6	
Within subjects		2062.3		
Trials	2	277.3	138.6	59.2*
Marks	2	11.7	5.8	2.5
Segments (P)	2	32.8	16.4	7.0*
Halves	1	68.1	68.1	29.3*
Relative segment (R) ^a	2	22.5	11.2	4.8*
MP _r	2	2.2	1.1	
HT	2	0.6	0.3	
HM	2	0.03	0.015	
HP	2	12.6	6.3	2.74
HR	2	6.3	3.1	
HMP	2	2.2	1.1	
Error	693	1626.0	2.34	
Total	755	5691		

^a Segments relative to mark.

**p* < .01.

between trials during the interval in which clock faces were changed. There is also a tendency for performance in each trial to be worse than in the one before, but this is less marked than in the previous experiment. The percentage of missed signals on the Blank face was so high in this experiment (see Figure 3) that the differences among groups of subjects with this condition in the successive trials introduced large variability.

EXPERIMENT III

Since the introduction of a single segment was found to decrease the percentage of missed signals to about half, this effect was studied in greater detail.

A preliminary experiment compared the effect of one mark of a $\frac{1}{2}$ -second width with that of a mark of 3 seconds width; no difference in effect was found. Three groups of seven subjects each were tested in three sessions on seven conditions, which included the two single marks in three different positions and also the Blank face. While 58.2% signals were missed from the Blank face, 21.8% were missed from the face with the $\frac{1}{2}$ -second mark, and 22.9% from the face

with the 3-second mark. There were no significant differences due to the positions of the marks.

Experiment III was designed to study the effect of proximity to the mark on the proportion of missed signals. If the mark were serving as a comparison point, it would be expected that fewer signals would be missed when the hand was near the mark.

Method

Three clock faces were used, each carrying one white mark of one $\frac{1}{2}$ -second width, in one of three positions, 12, 4, or 8 o'clock. Forty-two subjects were used, divided into three groups, each of which was tested on all three faces in one session, in a Latin square order. The signal was a pause of .4 second duration in the sweep of the white-tipped hand. Since seven subjects were tested at once, half the subjects served in a replication of the design.

Each subject was given two switches and was instructed to press one if she was sure that there was a signal and the other if she thought that there was the slightest chance that there might have been one. It was emphasized that it was very important not to miss any signals. It was hoped in this way to increase the proportion of false alarms in order to see if these changed in any consistent way, but this hope was not realized.

Sixty signals were given in a 10-minute trial, followed by a pause of a minute or so while the mark was moved to another position. The interval between signals ranged randomly from 5 to 14 seconds. Twenty signals were given in each of the three 20-second segments of the clock face centered around the three mark positions; i.e., around 12 o'clock the segment extends from 50 to 10 seconds past the minute, around 4 o'clock from 10 to 30, and around 8 o'clock from 30 to 50 seconds.

Results

The only important effect was the effect of order. Figure 2 shows the increase in the percentage of missed signals that occurred during the half-hour session of three trials. As before, as the trial progressed an increase was noted in the percentage of signals not reported. Table 1 shows the analysis of variance for the data. The difference between halves and the differences between trials were significant. Since the position of the mark did not significantly affect the number of missed signals, and there was no significant interaction between groups, trials, and positions, the curve probably is closely similar to that which would have been obtained if the mark had been kept in one position throughout.

The effect of the position of the signal also was significant (P in Table 1). Fewest signals (34.7%) were missed when they occurred in the top segment of the clock, centered around 12 o'clock, and most (39.7%) were missed in the segment around 8 o'clock. The entry "Relative segment" tests the hypothesis mentioned above that fewer signals would be missed when they were near the mark. The results showed that least signals were missed (34.7%) when the signal appeared in the same segment as the mark. The poorest performance was obtained (38.9% missed) in the segment preceding the mark.

Table 1 also shows a significant Trials \times Segments interaction. In the first trial the best performance appeared in the 4 o'clock segment; while in the two later trials, performance in the 12 o'clock segment was superior.

The proportion of "doubtful" responses was very low, as shown in Figure 2. These responses show the same trend as the missed signals; i.e., an increase within trials and from one trial to the next. False responses

were even fewer, with means of .5, .4, and .7 false responses per subject for the three trials. Within the trials the first half showed more false responses than the second. The data are too few to justify conclusions, and further experiments are underway to investigate this point more thoroughly.

Discussion

These experiments lead to the following conclusions. First, the proportion of signals detected is greatly increased by the addition of white marks to the black clock face. This improvement is not due to any change in the expected location of signals, for it is found even when the signals may occur anywhere on the clock face. It seems that an anchoring point makes it easier to detect a brief pause in the rotating hand. No difference has been found between the effects of one $\frac{1}{2}$ -second mark and two 3-second marks, but with four 3-second marks an increased effect is apparent.

Second, a decrement in detection throughout each session is observed even though the signals are given at a mean rate of one per 10 seconds. This decrement occurs even within a 6-minute trial. There is a tendency toward improvement during the pause while the clock face is changed, but there is also a decrement from trial to trial during the first 30 minutes of the session. Thus the form of the decrement is similar to that found in ordinary vigilance tasks, where signals are few and highly irregular and the general situation is much more monotonous. A number of authors (e.g., McFarland, Holway, & Hurvich, 1942; Zwislocki, Maire, Feldman, & Rubin, 1958) have shown that deterioration occurs with continuous measurements of thresholds. Haider and Dixon (1961) reported a continuous increase in the brightness of a spot which the subject was required to keep just visible. This change occurred over the first half hour, after which no further change occurred. They suggested that continuous measurements of thresholds could be used as sensory indicators of vigilance. It is clear that infrequency of signals is not essential to produce a decrement in detection of signals in a vigilance experiment.

Experiment II confirmed the previously re-

ported finding that, when the number of white marks was kept constant, more signals were detected when they were restricted to the black segments than when they could occur anywhere on the clock face. This is in line with the expectancy theory, which states that the probability of detection of a signal is related to the probability of its occurrence at that point.

Finally, the results from Experiment III indicate that signals are detected more easily when they occur near a fixed reference mark. It is perhaps surprising that this proximity effect is so small, when the overall effect of the mark on detection is so large.

REFERENCES

- HAIDER, M., & DIXON, N. F. Influences of training and fatigue on the continuous recording of a visual differential threshold. *Brit. J. Psychol.*, 1961, **52**, 227.
- McFARLAND, R. A., HOLWAY, A. N., & HURVICH, L. M. *Studies of visual fatigue*. Cambridge: Harvard Graduate School of Business Administration, 1942.
- MACKWORTH, J. F. The effect of intermittent signal probability on vigilance. *J. Canad. Psychol.*, 1963, **17**, 82-89.
- ZWISLOCKI, J., MAIRE, F., FELDMAN, A. S., & RUBIN, H. On the effect of practice and motivation on the threshold of audibility. *J. Acoust. Soc. Amer.*, 1958, **30**, 254.

(Received July 12, 1962)

SERIES EFFECTS IN MOTOR PERFORMANCE STUDIES

J. E. KENNEDY AND J. LANDESMAN

Bureau of Industrial Psychology, University of Wisconsin

It was hypothesized that a series effect would operate in a study designed to determine the optimal work surface height for the performance of a simple motor task. Under Condition A, Ss performed the task at each of 6 work surface heights, comprising the lower $\frac{2}{3}$ of the range of heights used in an earlier study. Under Condition B, Ss performed at 6 heights comprising the upper $\frac{2}{3}$ of the range. Systematic differences were observed in the performance of the 2 groups at the 4 heights they had in common. The differences in motor performance were attributed to a series effect stemming from differences in judgments Ss made concerning what the optimal height should be.

A number of studies reported in the human engineering literature might be characterized as having the general purpose of determining the optimal level of some stimulus condition for the performance of a particular task. Examples of the independent variables investigated in studies of this kind are differences in radii of handcranks (Reed, 1948), distance of a foot control from the back of a seat (Gough & Beard, 1936), and the angle of a display in remote handling equipment (Baker, 1960). The dependent variable is typically some measure of speed or accuracy of performance on a particular task.

The experimental designs used in these studies are of two general types. In one type, a subject performs the task at only a single stimulus level; in the other type, a subject performs the task at each of the several stimulus levels being investigated. For both of these designs it is not uncommon for the results to take the general form of an inverted U. That is, the optimal stimulus level is found to occur at or near the center of the range of stimulus levels used and performance falls off as the extremes are approached in either direction.

Our concern is with the possibility of a series effect operating in human engineering studies of this kind in which the subject performs at several stimulus levels. It has long been known that such series effects operate in the realm of the psychology of judgment. As a single example, Kennedy (1961) instructed subjects to make esthetic judgments of isosceles triangles of varying

proportions and found that the most preferred proportion was at or near the center of the range used. A series effect was inferred to be operating because when the same stimulus is included in a different series its preference value shifts considerably, depending on whether its position is near the center of the range or near the extremes. Furthermore, efforts to obscure the range of stimuli under consideration by using a paired comparison technique in which the stimuli are only presented two at a time in random order, does not obliterate the series effect.

If it should be found that such a series effect operates in human engineering studies concerned with, for example, motor performance, it would render equivocal the conclusion as to what is *the* optimal stimulus level for it would depend upon the particular series of stimulus levels used.

GENERAL PURPOSE OF THE STUDY

A study was selected from the literature of human engineering concerned with the problem of determining the optimal stimulus level for performing a motor task, which employed the type of design in which each subject performed at all stimulus levels, and which yielded results in the general form of the inverted U. We replicated the experiment as closely as possible with one major change. One half of our subjects performed the task at each of the stimulus levels comprising the *lower* two-thirds of the range used in the original study; the other half of our subjects performed the task at each of the

stimulus levels comprising the *upper* two-thirds of the original range.

If no series effect operates under these conditions we would not expect the optimal stimulus levels for the two groups to differ by any amount greater than could be accounted for by sampling error. On the other hand, if a series effect does operate we would predict the optimal stimulus level to shift toward the center of the range used for each condition. Explicitly, we hypothesized that two curves, generally parabolic in form, would be obtained, and that the stimulus level associated with maximum performance would be significantly different.

PROCEDURE

We readily found a study meeting our requirements. Ellis (1951) investigated the optimal work surface height for the performance of a simple motor task (a modified version of the Minnesota Rate of Manipulation Test). The work surface height was adjustable and each subject performed the task at six heights ranging from 18.9 inches below the elbow to 7.9 inches above. The independent variable was work surface height and the dependent variable was the number of blocks turned on the manipulation test during a 3-minute trial. A 2-minute rest period was introduced between trials. During this interval the subject was given knowledge of results, completed a rating scale concerned with the degree and locus of fatigue associated with the trial, and relaxed for the remaining time. Ellis found that when mean performance y was plotted against work surface height x , y was found to be a nearly symmetrical, decreasing function of x about the center of the range of heights employed.

Our sample of 36 subjects (volunteer students

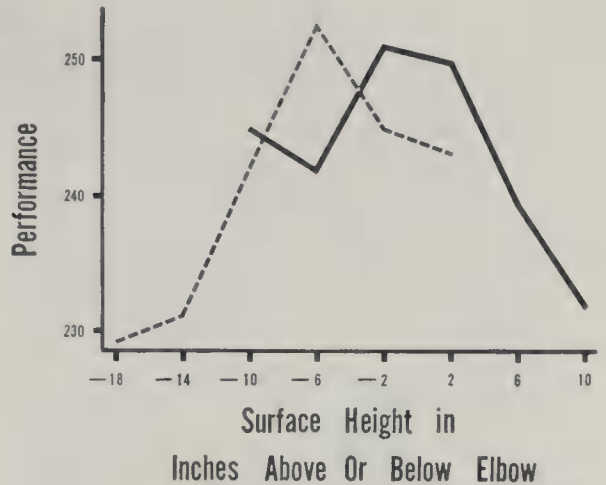


FIG. 1. Mean task performance as a function of work surface height for Condition A—broken line, $N = 18$ —and Condition B—solid line, $N = 18$.

from an introductory psychology course) were randomly assigned to one of six 6×6 Latin squares (Order \times Work Surface Height). One half of the subjects (Condition A) performed at each of the following six work surface heights: -18 , -14 , -10 , -6 , -2 , and $+2$ inches relative to elbow height. The remaining one half (Condition B) performed at -10 , -6 , -2 , $+2$, $+6$, and $+10$ inches relative to elbow height. Note that while two heights were unique for each condition, four heights (-10 , -6 , -2 , and $+2$) were common to the two conditions.

RESULTS

The results are presented graphically in Figure 1 and appear to be consistent with the hypothesis. The optimal stimulus level seems to differ for the two conditions and in the expected manner. For descriptive purposes, parabolas were fitted to each of the

TABLE 1
SUMMARY OF ANALYSES OF VARIANCE FOR CONDITIONS A AND B

Source	Condition A			Condition B	
	<i>df</i>	<i>MS</i>	<i>F</i>	<i>MS</i>	<i>F</i>
Squares	2	6747.81		885.86	
Between Ss (same square)	15	1227.10		2090.45	
Heights	5	1375.85	4.44*	918.68	8.04**
Trials (order)	5	1007.79		1672.39	
Squares \times Trials	10	111.96		52.07	
Squares \times Heights	10	309.99	2.41*	179.16	1.57
Error	60	128.47		114.31	
Total	107				

* $p < .05$.

** $p < .01$.

TABLE 2
SUMMARY OF TREND TESTS FOR CONDITIONS A AND B

Source	Condition A			Condition B		
	df	MS	F	df	MS	F
Linear component	1	3641.40	11.75**	1	1422.95	12.45**
Quadratic component	1	1798.41	5.80*	1	2367.50	20.71**
Error	10			60	114.31	

* $p < .05$.
** $p < .01$.

two curves yielding the following equations for Condition A and Condition B, respectively: $y = 211.19 + 16.00x - 1.77x^2$ and $y = 205.36 + 17.78x - 1.81x^2$. The first derivative of these equations gave maxima at x values of -3.92 inches for Condition A and -2.34 inches for Condition B; a difference of 1.58 inches.

Whether or not we can consider these results as statistically significant must be considered next.

An analysis of variance was employed separately for Condition A and Condition B data to establish that the six means differed significantly from one another. The results of these analyses are found in Table 1. It should be noted that under Condition A the Squares \times Heights interaction term reached significance at the .05 level and hence became the appropriate error term for testing differences among heights. Since this was not the case in Condition B, the residual was used.

TABLE 3
SUMMARY OF ANALYSIS OF VARIANCE
BETWEEN EXPERIMENTAL CONDITIONS

Source	df	MS	F
Conditions	2	49.00	
Error (a)	34	1412.70	
Heights	3	130.17	
Conditions \times Heights	3	599.57	3.26*
Error (b)	102	183.64	
Total	143		

* $p < .05$.

The analysis of variance was extended then to trend tests to test the hypothesis that the quadratic component for each experimental condition was significant. The results are summarized in Table 2 and support the hypothesis and, incidentally, are consistent with those obtained in the original study by Ellis.

The remaining step was to demonstrate that the two curves are not parallel across the four heights common to the two experimental conditions; i.e., that the sum of squares for the interaction of Conditions \times Heights departed significantly from zero. When the analysis is limited to these four heights the Latin squares are no longer complete and learning effects, known to be a major source of variance, were left in the error term. Despite this invitation to a Type II error, the interaction reaches the .05 level as can be seen in the summary of this analysis in Table 3.

The implications of the analysis become clear if one notes that when the data from the two experimental conditions are pooled, which amounts to counterbalancing for series effects, the differences in mean performances across heights are not significant (Table 3) and are, in fact, surprisingly similar (243.6, 247.3, 247.6, and 246.5). In other words, there was an optimal height under Condition A and under Condition B but not under Conditions A and B combined.

DISCUSSION AND CONCLUSIONS

The experimental design employed in this study was aimed at determining whether or not a series effect occurs and, as such, provides little or no analytical data to indicate why it should occur. The similarity between

these results and those found in judgment studies suggests the possibility that the relationship between work surface height and task performance is not simple and direct but rather one which is moderated by the judgments or expectancies of the subjects. We suspect that all subjects do not maintain a constant effort at each of the six heights but rather make an extra effort at those heights which they judge to be more comfortable and which, from their common sense viewpoint, are judged to be near the optimal level. Such judgments would necessarily be crude since, at the outset of the experiment, the subject knows he will work at six heights but does not know what they will be. His judgment of any given height would be expected to be influenced by the particular series of heights in which it appears, and may invite an error of central tendency; i.e., the tendency for stimuli to be judged in the direction of the central region or point in the series (Guilford, 1954). This kind of explanation would seem to fit the facts. The subjects perform systematically differently at the same heights when they appear in different series because they make systematically different judgments of these heights as they appear in different series.

The implications of these results are clear. If only a single series of stimuli are used in studies of this kind the optimal stimulus level must be considered optimal relative to the series employed, not optimal in any absolute sense. Use of an experimental design

in which each subject operates at only a single stimulus level would eliminate a series effect. Whether such a procedure would justify the larger number of subjects required would depend on the specific purpose of the experiment. It would seem to be a necessary safeguard if we have reason to believe that the true optimal level might be markedly different from that which subjects might anticipate it to be and it is in this kind of situation that applied experimental psychologists often suggest they might make their most telling contribution.

REFERENCES

- BAKER, D. F. Task performance with the CRL Model 8 master-slave manipulator as a function of object size, angle, and height of display. *USAF WADC tech. Rep.*, 1960, No. 60-167.
- ELLIS, D. S. Speed of manipulative performance as a function of work surface heights. *J. appl. Psychol.*, 1951, **35**, 238-296.
- GOUGH, M. N., & BEARD, A. P. Limitations of the pilot in applying forces to airplane controls. Technical Notes No. 550, 1936, National Advisory Committee on Aeronautics. Cited by E. J. McCormick, *Human engineering*. New York: McGraw-Hill, 1957.
- GUILFORD, J. P. *Psychometric methods*. (2nd ed.) New York: McGraw-Hill, 1954.
- KENNEDY, J. E. The paired-comparison method and central tendency effect in esthetic judgments. *J. appl. Psychol.*, 1961, **45**, 128-129.
- REED, J. D. Factors influencing rotary performance. Unpublished doctoral dissertation, Johns Hopkins University, 1948. Cited by A. Chapanis, W. R. Garner, and C. T. Morgan, *Applied experimental psychology*. New York: Wiley, 1949. P. 286.

(Received July 16, 1962)

ASSESSMENTS OF INNOVATIVE BEHAVIOR: PARTIAL CRITERIA FOR THE ASSESSMENT OF EXECUTIVE PERFORMANCE¹

GARLIE A. FOREHAND²

Center for Programs in Government Administration, University of Chicago

Assessments by superiors and peers of the innovative behavior of administrators are considered as partial criteria of executive performance. A measure based upon 7-point rating scales was adjudged unacceptable as a measure of innovative behavior, because of insufficient discriminant validity with respect to ratings of other attributes, and because of uniformly high correlation with ratings of general effectiveness, regardless of raters' independently expressed attitudes toward innovative behavior. A measure based upon forced choice between innovative and noninnovative descriptions shows more promise: its major correlates both within and across raters are other measures of innovation and of attributes theoretically related to innovativeness; and it is significantly correlated with general effectiveness ratings only when assessors report, by an independent measure, that they value innovative behavior highly. Assessments are influenced by status of assessor (supervisor or peer) and by organizational climates.

A major difficulty in conducting research on executive performance has been summarized by Thorndike (1961):

The criterion problem . . . represents the heart of the problem of doing research on executive appraisal. Getting suitable criterion measures for executives is almost uniquely difficult because the diversity of the job and the intangibility of its products make it almost impossible (1) to get any tangible performance indications of job success, and (2) to get comparable data for different individuals (pp. 69-70).

One suggested approach to the criterion problem is to focus upon *partial criteria* as opposed to a global or unitary criterion—upon “the assessment . . . of particular facets of executive behavior that are regarded . . . as distinguishable components of executive activities” (Tagiuri, 1961, p. 110). Such an approach divides “the criterion problem” into two components: the assessment of particular behaviors, and determination of the relevance of the selected behaviors to the overall success or value of the executive. The distinction is similar to that made in a decision-theory approach between the “outcome” and “payoff” matrices (Cronbach & Gleser, 1957).

¹ This study was conducted as part of Cooperative Research Project No. 975, supported by the U. S. Office of Education. Appreciation is expressed to Harold Guetzkow for many contributions to the study and this paper.

² Now at Carnegie Institute of Technology.

Each component presents questions for empirical research. Obtaining measures of “particular facets of behavior” is not an easy matter, particularly since such measures are almost necessarily based upon the judgment of observers, with attendant problems related to interpersonal perception (Gage & Cronbach, 1955). Once a partial criterion is defined, its “practicality” and its relationship to overall effectiveness both depend upon aspects of the situation in which the study is made (Guetzkow & Forehand, 1961).

These issues provide context for a study of assessments by superiors and peers of innovative behavior in government administrators, designed for use as criteria in a predictor study and as measures for evaluating an educational program. Innovative behavior as the term is used here includes the development and consideration of novel solutions to administrative problems, and evaluation of them in terms of criteria broader than conformity to pre-existing practice. The point of view taken here is that the measures obtained are reports of interpersonal perception, and are therefore influenced by the instruments used, by characteristics of the assessors, and by the nature of the organizations in which the assessments were made. Empirical questions relating to each of these influences may be outlined briefly.

1. *Validity of Trait Measures*—If an instrument is “valid” to the extent that it “measures what it is supposed to measure,” then questions of validity are clearly applicable to partial criterion measures. Some questions regarding validity of measures of innovative behavior concern “face validity,” and are not easily subjected to empirical analysis. Some important questions regarding discriminant and convergent validity, however, need to be investigated (Campbell & Fiske, 1959). Is innovative behavior, as attributed to an executive by a rater, discriminable from other executive traits (e.g., analytic problem solving abilities or policy executing skills), or from general effectiveness? Does one rater’s perception of innovative behavior correspond to another rater’s perception of that variable, more than to the other rater’s perception of other traits? If the measures are to be useful, we would expect two innovation measures as provided by a particular set of raters (e.g., superiors or peers) to be more highly correlated with each other than either of them would be with any other variable. We also would expect the correlation between peers’ and superiors’ assessments of innovative behavior to be higher than the correlation between an innovation score obtained from one set of raters and other scores obtained from the other set of raters.

2. *Effect of Raters’ Attitudes*—It is clear that ratings are influenced by characteristics of the raters as well as by characteristics of the persons rated (Bruner, 1951, Tannenbaum, Weschler, & Massarik, 1961, pp. 324–332). With respect to innovative behavior as a partial criterion, a rater characteristic of special relevance is the rater’s attitude toward innovative behavior which may be expected to influence both sensitivity to the behavior and its relation to perceived success. Attitudes of raters may, for example, influence “interrater reliability” of ratings. If two raters tend to agree in the values they attach to innovative behavior, they might also be expected to agree in their general evaluations of an individual’s performance. Raters’ evaluation of innovative behavior should be directly related to raters’ attitudes toward innovation. In the present study we may consider that (a) directly assessed attitude to-

ward innovation and (b) the degree to which persons seen as innovative are also seen as generally effective are separate measures of the same variable. It is predicted that the correlation between innovation scores and global ratings will be higher when raters report that they value innovation highly than when they do not so report.

3. *Effect of Perceived Organization Climate*—“Climate” is another aspect of the “situation” which may determine whether or not a given kind of behavior leads to perceived success. Organizational practices which can be described as “bureaucratic” or rule centered appear to be those that would discourage innovative behavior on the part of individual administrators (March & Simon, 1958, pp. 36–47), while “democratic” or group centered practices ought to encourage such behavior. It is predicted, therefore, that the correlations between innovation scores and global ratings will be higher in organizations that subjects describe as democratic than in organizations they describe as bureaucratic.

PROCEDURE

Variables

The variables³ discussed in this paper are derived from questionnaires by means of which government executives were described by organizational associates. Two groups of raters (superiors and peers) and two methods for describing behavior (rating scales and forced-adjective comparisons) were used. The scores examined are described briefly below.

All of the items described as “ratings” consist of brief complimentary statements “which may be descriptive of government executives.” The rater responds on a seven-point scale, the alternatives ranging from “perfectly describes the participant” to “could not be accurately applied to the participant.” The following 10 items, which were distributed throughout a 60-item questionnaire, constitute the Innovation Rating Cluster (Variable 1):

³ The following materials have been deposited with the American Documentation Institute: A copy of the questionnaire containing the rating scales, the forced-choice assessment, and the raters’ attitude measure; a copy of the Organizational Perceptions instrument; and a four page paper reporting several supplementary analyses. Order Document No. 7503 from Auxiliary Publications Project, Photoduplication Service, Library of Congress, Washington 25, D. C. Remit in advance \$2.00 for microfilm or \$3.75 for photocopies and make checks payable to: Chief, Photoduplication Service, Library of Congress.

1. Approaches problems with flexibility; able to change assumptions and orientations in order to arrive at workable new decisions.

2. When decisions require inventiveness: maintains flexibility in developing a wide variety of approaches to a decision.

3. Works creatively to develop new and unusual solutions to the organization's problems.

4. Is quick to recognize when events require decisions to re-evaluate organizational practices and implications.

5. Constantly evaluates his own methods and practices, adapting them to changing conditions when necessary.

6. When a decision cannot repeat past solutions or be modeled after precedent: probes for fresh new orientations before coming to final judgment.

7. When a situation demands innovation: relies upon his own ideas and judgment.

8. Able to free himself from binding considerations in coming to independent judgments.

9. When decisions involve new and unsolved problems: is creative and imaginative in developing new solutions.

10. Able to develop and implement independent, innovative, and original judgments.

In the split-half analysis reported below, the 10 items were divided randomly into two halves, presented here as Items 1-5 and 6-10, respectively (thus, the order presented here is not that in which the items were marked).

The Forced-Choice Innovation Measure (Variable 2) was constructed from a list of five adjectives judged to be descriptive of innovative behavior (self-reliant, inquiring, flexible, original, independent) and five adjectives judged to be equally complimentary but unrelated to innovation (stable, industrious, prudent, dependable, cooperative). The 25 "cross list" pairs of adjectives, i.e., pairs containing one adjective from each list (along with 10 "dummy" intralist pairs), constituted the material presented to the rater. The rater was asked to "select and underline the one adjective from each pair which you think is more descriptive of the participant." The score is the number of times an "innovative" adjective was selected over a "noninnovative" adjective.

The seven rating clusters pertaining to other facets of executive behavior, numbered 3 through 9, are each composed of five rating scale items. Their content and titles are based upon cluster analysis of a previous version of the scales. They are similar in content to similarly named scales described elsewhere (Forehand & Guetzkow, 1962; Guetzkow, Forehand, & James, 1962). Although they may be expected to correlate to varying degrees with innovation ratings, each is conceptually distinguishable from innovation and ought to be distinguishable empirically.

The Global Evaluation Rating Cluster (Variable 10) consists of three items:

1. Capable of handling almost any executive-level job successfully

2. In general, is an extremely competent manager

3. One of the best administrators I have ever known

This cluster was included in an attempt to define a separate "halo" or general evaluation measure.

The Rater's Attitude Variable (Variable 11) utilized the same adjective pairs as Variable 2 (in different order), but raters were instructed to describe "your conception of an *ideal government executive*," rather than a particular individual. The rater's attitude-toward-innovation score is the number of innovative adjectives chosen, high score interpreted as indicating a favorable attitude toward innovation.

Perceptions of organizational climate were assessed by means of an adaptation of Nelson's Leadership Practices Survey (Nelson, 1959). The instrument is designed to assess attitudes toward leadership. Fifty administrative problems or situations are described, and the subject selects one of two alternative responses to each of them. The responses are keyed as representing differing orientations toward leadership: bureaucratic or rule centered, technocratic or work centered, idiocratic or individual centered, and democratic or group centered. In the present study, subjects were asked to indicate the practices which they would consider typical in their organizations, rather than their own preferences; the wording of items and instructions was modified to facilitate such responses. The alternatives are presented in pairs. In the form used here rule centered alternatives are always paired with group centered ones, so that rule centered and group centered descriptions are treated as if they were on a continuum. The 25 items involving these two kinds of alternatives were used in the analysis.

Subjects and Raters

The subjects whose behavior was described by the raters were 188 persons holding administrative positions in 30 agencies of the United States Government. Their grade levels range from 11 to 17, with a mean of 13.3. Eighty-seven described their positions as "primarily line" and 101 described their positions as "primarily staff" positions. The subjects responded to the organizational perception questionnaire in group sessions during which they also took other tests for use in a later predictor study. Each subject was described by his immediate superior and by a peer (identified as a "co-worker who knows you well and who is neither a superior nor a subordinate") who was nominated by the subject himself. Superior and peer questionnaires were distributed and collected by mail.

RESULTS

Validity of Trait Measures

The 10 items in the Innovation Rating Cluster (Variable 1) were included together in a single scale on the basis of a prior judg-

ment rather than predetermined empirical homogeneity. Their internal consistency should exceed the relationship of that cluster to other variables, if the theoretical convergence in their interpretation is mirrored in empirical convergence when they are used to describe individuals. The uncorrected correlation between the two halves of the Innovation Rating Cluster (Variables 1) is .87 for superiors' ratings and .84 for peers' ratings. Correlations between these half-cluster scores and other rating clusters range from .52 to .78 for superiors, and from .25 to .74 for peers. In each case the correlation between the two half-cluster innovation scores is significantly ($p < .05$) greater than all of the correlations between an innovation half-cluster and other clusters. We may thus conclude that the innovation cluster has homogeneity independent of its relationship with other rated attributes, even though there is evidence of considerable generalized "halo" variance.

Results concerning intrarater relationships among measures are presented in Table 1. The prediction that an innovation measure would be more highly correlated with another innovation measure than with other variables is not supported with reference to the Innovation Rating Cluster (Variables 1). In both superiors' and peers' ratings the correlation

between Variables 1 and 2 is almost invariably *smaller* than those between Variable 1 and other rating scales, in most cases substantially so. "Partialing" these correlations with respect to the Global Evaluation Rating Cluster in an attempt to control halo variance, results in a smaller discrepancy between innovation-innovation and innovation-other trait correlations, but the latter are still, in general, substantially higher than the former. The rating cluster scores evidently share "method factor" variance to such an extent that different traits measured by the same method correlated more highly than do different measures of the same trait.

The Forced-Choice Innovation Measure (Variable 2) more closely manifests the expected properties. With both sets of raters, the correlation between Variables 1 and 2 significantly exceeds all but one of the seven correlations between Variable 2 and measures of attributes other than innovation ($p < .05$). The exception in each case is the Self-Confidence Rating Cluster. In terms of partial correlations, correcting again for relationships with the Global Evaluation Rating Cluster, the only variables that have a nonzero relationship with the Forced-Choice Innovation Measure are the rating clusters for Innovation, Self-Confidence, and (inversely) Cautiousness (Variables 1, 3, and 4). The low but

TABLE 1
INTRARATER CORRELATIONS

Variable	Zero-order correlations				Partial correlations ^a			
	Superior		Peer		Superior		Peer	
	Var. 1	Var. 2	Var. 1	Var. 2	Var. 1	Var. 2	Var. 1	Var. 2
1. Innovation Rating Cluster	(.93) ^b	.44 ^c	(.92) ^b	.37 ^c	—	.36 ^c	—	.32 ^c
2. Forced-Choice Innovation Measure	.44	—	.37	—	.36	—	.32	—
3. Self-Confidence Rating Cluster	.79	.43	.70	.38	.59	.48	.41	.33
4. Cautiousness Rating Cluster	.48	-.18	.31	-.25	.28	-.48	.26	-.31
5. Discernment Rating Cluster	.79	.22	.70	.12	.59	.05	.43	-.02
6. Policy Making Rating Cluster	.80	.25	.73	.22	.60	.1	.50	.10
7. Policy Execution Rating Cluster	.70	.20	.55	-.02	.45	.04	.26	-.18
8. Bureaucratic Decision Making Cluster	.77	.20	.72	.12	.54	.05	.55	.00
9. Analytic Decision Making Rating Cluster	.75	.20	.74	.10	.51	.06	.49	-.06
10. Global Evaluation Rating Cluster	.73	.29	.74	.21	—	—	—	—

Note.— $N = 188$; $r = .14$, $p < .05$; $r = .19$, $p < .01$.
^a Controlling variance attributable to correlation with Variable 10.
^b Corrected split-half reliability.
^c Entry repeated from preceding column.

TABLE 2
INTERRATER CORRELATIONS OF INNOVATION MEASURES

	Variable	
	1	2
Interrater "reliability" (uncorrected): r_{II}	.22**	.32**
Summary of interrater correlations: Innovation measures with Variables 3-9		
Median correlation ($N = 14$)	.18	.08
Highest	.27	.22
Lowest	.02	.01
Number higher than r_{II}	1	0
Number not significantly lower than r_{II} *	8	1

Note.— $N = 188$.
* $p < .05$ (one-tailed test).
** $p < .01$.

significant correlations between Variables 2 and 10 suggest, further, that innovative behavior, as manifested in forced-adjective comparisons, can be perceived relatively independently of general effectiveness, even though it is perceived as positively related to general effectiveness.

Interrater (superior versus peer) reliabilities for both innovation measures are significantly greater than zero, as is shown in Table 2. The interrater reliability for the Innovation Rating Cluster numerically exceeds all but one of the 14 cross-variable interrater correlations, but significantly exceeds only 6 of them. The correlation between two sets of responses to the Forced-Choice Innovation Measure numerically exceeds all of the other cross-rater correlations involving that measure, and significantly exceeds all except one ($p < .05$). The exception: Innovation as reported by peers on Variable 2 has a correlation of .22 with Self-Confidence (Variable 3) as reported by superiors. The predictions regarding interrater relationships are thus supported, in general, for Variable 2, but are not supported for Variable 1.

Effect of Raters' Attitudes

An analysis of the relationship between degree of agreement between raters on the attitude measure and interrater correlations on several assessment scales yielded ambiguous results. A description of the analysis, results, and tentative interpretations has been made available elsewhere (see Footnote 3).

The effect of raters' attitudes upon evaluation of innovative behavior was examined by comparing the correlation between innovation measures and general effectiveness ratings in (a) the 60 subjects whose superiors had the highest attitude-toward-innovation scores and the 60 whose superiors had the lowest attitude scores, and (b) the 60 subjects whose peers had the highest attitude scores and the 60 whose peers had the lowest attitude scores. The results, as shown in Table 3, indicate that, as predicted, the Forced-Choice Innovation Measure has a significantly higher correlation with the general effectiveness score when raters report a favorable attitude toward innovation in the abstract. It will also be noted that the correlations between these two measures when raters favor innovation (.51 and .46 for superiors and peers, respectively) are substantially greater than the corresponding values in the total sample (.29 and .21; see Table 1); the significance of these differences was not tested since the smaller samples are subgroups of the larger. Thus rater attitude toward innovation is convergently manifested in direct measurement and in the valuation of individuals perceived to vary in innovative behavior. These relationships are not replicated with respect to the Innovation Rating Cluster. It appears that the latter measure is so highly saturated with generalized rater variance that it is not possible to observe a differential between relationships of this measure to general effectiveness ratings as a function of rater attitude.

TABLE 3

INTRARATER CORRELATIONS WITH THE GENERAL EFFECTIVENESS RATING IN SUBSAMPLES
DEFINED BY RATERS' ATTITUDES TOWARD INNOVATIVE BEHAVIOR

Variable	Superiors		Peers	
	Favorable attitude	Unfavorable attitude	Favorable attitude	Unfavorable attitude
1. Innovation Rating Cluster	.74	.66	.82	.75
2. Forced-Choice Innovation Measure	.51	.12**	.46	.02**

Note.— $N = 60$.

** Significantly lower than value in preceding column ($p < .01$).

Effect of Perceived Organizational Climate

The influence of organizational climate upon evaluation of innovative behavior was studied by examining results separately for the 60 subjects who most often described their organization's practices as group centered and the 60 subjects who most often described their organization's practices as rule centered. Table 4 shows that for peer responses and for combined superior and peer responses (sum of standard scores), the Forced-Choice Innovation (Variable 2) is significantly correlated with Global ratings in organizations described as group centered but not in organizations described as rule centered. The direction is as predicted, but the difference between the correlations in the two samples is not significant. With respect to superiors' ratings, the relationships are reversed. Variable 2, as responded to by superiors, is significantly related to Global ratings in the rule centered subsample but not in the group centered subsample. Again, the cross-sample differences are not significant. Correlations of Variable 1 with Global ratings are numerically but not significantly

higher in the group centered than in the rule centered subsamples.

DISCUSSION

It is evident, both from the results of the present study and from those of previous studies (Fiske & Cox, 1960; Forehand & Guetzkow, 1962; Mayo, 1956), that lack of discriminant validity of ratings is a serious obstacle to the study of partial criteria. The rating scale measure of innovative behavior employed in this study (and incidentally, those of other traits) can be adjudged unacceptable as an index of a particular, distinguishable facet of behavior, because of inability of raters to differentiate among the behavior patterns as defined. On the other hand, the Forced-Choice Innovation Measure—with which observers select from adjective pairs those adjectives which seem most descriptive of a subject—seems more promising. The correlation between superiors' and peers' responses to the variable is sufficient to indicate that the raters' perceptions of the behavior it describes are to some degree congruent: its major correlates, both within and

TABLE 4

INTRARATER CORRELATIONS WITH THE GENERAL EFFECTIVENESS MEASURE IN SUBSAMPLES WHOSE MEMBERS DESCRIBED THEIR ORGANIZATIONS AS GROUP CENTERED (GC) OR RULE CENTERED (RC)^a

Variable	Superiors' ratings		Peers' ratings		Combined ratings ^a	
	GC	RC	GC	RC	GC	RC
1. Innovation Rating Cluster	.70*	.67*	.77*	.63*	.74*	.62*
2. Forced-Choice Innovation Measure	.12	.27*	.35*	.14	.31*	.19

Note.— $N = 60$.

^a Superiors' + peers' innovation assessment versus superiors' + peers' general effectiveness assessment.

* $p < .01$.

across raters, are other measures of innovation, and ratings of attributes conceptually and theoretically associated with innovativeness; it is not highly related to general effectiveness ratings except when raters indicate by an independent measure that they place a high value on innovative behavior in general. To illustrate the suggested properties of the two measures: if objectively measured correlates of the two scores are studied, we may predict that the variables that would be most closely related to the innovative rating cluster would be those that would be expected to predict general success, e.g., general mental ability, achievement motivation; the variables most closely related to the forced-choice innovative score may be interpreted more specifically as attributes related to innovative behavior. The results offer encouragement that further work on forced-choice rating methods (e.g., Kay, 1959) might provide discriminable partial criteria, particularly if some attention is given to the comparative social desirability (Edwards, 1959) of the paired descriptions.

Some limitations in the analysis of convergent and discriminant validity in the present study should be noted. First, it is clear from a comparison of Tables 1 and 2 that two raters' assessments of the same trait are not more highly correlated than are two traits as assessed by the same rater, as the Campbell-Fiske criterion would require; the studies of ratings cited above would suggest that such a relationship could not be reasonably expected. It should be noted, however, that measures remain both theoretically interesting and potentially useful when there is evidence that variance ascribable to content exists over and above that attributable to rater bias (Humphreys, 1960). Secondly, while rating cluster responses were compared with similarly defined measures of other traits, only the trait of innovative behavior was assessed by the forced-choice method. The omission—which resulted from failure to anticipate the apparent superiority of the Forced-Choice Innovation Measure—would need to be remedied in future studies of the validity of forced-choice measures.

Several results suggest interacting relationships between perceived innovative behavior,

the attitude and hierarchical perspective of the perceiver, and organizational climate. Perhaps these relationships may be summarized most briefly by stating conditions under which the value assigned to innovative behavior—and hence the utility of innovative behavior as a partial criterion—will vary. To an extent it is tautological to state that the general performance of persons perceived as innovative will be evaluated more highly by perceivers who themselves value innovative behavior highly than by those who do not. More pertinent here is evidence that the perceiver's attitude can be meaningfully assessed and can thus be incorporated into research designs and applied generalizations, as has been previously suggested (Guetzkow & Forehand, 1961). Also noteworthy is the fact that such a relationship was *not* manifested with respect to the rating scale measure. This indicates that empirical conformity to logical criteria of relationship between specific behaviors and general evaluations is an aspect of the validity of partial criterion measures that may not be neglected.

Results suggest that persons perceived as innovative will be evaluated more highly by peers if the organizational climate is perceived as group centered or "democratic," while they will be more highly evaluated by superiors if the climate is perceived as rule centered or "bureaucratic." In interpreting this result, it should be noted that the organizational levels of both assessors and assessee varied in the sample. The observation, if it proves to be replicable, must be attributed to the *relationship* between assessor and assessee, rather than to their absolute hierarchical positions. Blau and Scott (1962, pp. 116–128), in interpreting experimental studies of communication patterns, developed the hypothesis that a hierarchical relationship among group members (as between superiors and subordinates) facilitates *coordination* of group members, by restricting the flow of communication, while nonhierarchical relationships (as between peers) facilitates the development of *new solutions* to problems. Our findings are consistent with this hypothesis. In group centered environments, peers cooperate in the development of new problem solutions, and a superior is likely

to receive for review a group developed solution. In such a situation, the innovative skills of an individual would be valued by peers for their contribution to the group's task, whereas a superior would not be in a position to observe that contribution. In a rule centered organization, the contributions of peers are more likely to be individual in nature leaving the task of coordinating individual contributions to the superior. In this situation an administrator's innovative skills would tend to be irrelevant to the individual goals of his peers, but apparent and useful to the superior in achieving a coordinated group product. This analysis would suggest, as a hypothesis for future research, that innovative behavior would be more readily recognizable, and hence more predictable by cognitive and personality characteristics, when assessed by peers in group centered climates and when assessed by superiors in rule centered climates.

In summary, the factors that influence partial criterion measures of administrative performance are many and complex, particularly when measures take the form of ratings. Two aspects of such measures have been identified as requiring empirical attention: the distinguishability of the facets of performance examined and the ways in which the facets of behavior are evaluated. It may be suggested that partial criteria are likely to be "practical" only if such properties of the measures are examined under varied conditions. The results of the present study suggest that at least some relevant conditions are amenable to specification and measurement, although procedures for doing this are in an elementary stage of development. Such studies offer promise of contributing to theoretical understanding of behavior-in-organization and eventually an opportunity for improving upon what Henry (1961) has called "coefficients of the 'Irish' variety (O'Toole, O'Rourke, O'Seven, or .00)," at least under certain identifiable conditions.

REFERENCES

- BLAU, P. M., & SCOTT, N. R. *Formal organizations*. San Francisco: Chandler, 1962.
- BRUNER, J. S. Personality dynamics and the process of perceiving. In R. R. Blake & G. V. Ramsey (Eds.), *Perception: An approach to personality*. New York: Ronald, 1951.
- CAMPBELL, D. T., & FISKE, D. W. Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychol. Bull.*, 1959, **56**, 81-105.
- CRONBACH, L. J., & GLESER, G. *Psychological tests and personnel decisions*. Urbana: Univer. Illinois Press, 1957.
- EDWARDS, A. S. Social desirability and personality test construction. In B. M. Bass & I. Berg (Eds.), *Objective approaches to personality assessment*. New York: Van Nostrand, 1959.
- FISKE, D. W., & COX, J. A. The consistency of ratings by peers. *J. appl. Psychol.*, 1960, **44**, 11-17.
- FOREHAND, G. A., & GUETZKOW, H. Judgment and decision-making activities of government executives as described by superiors and co-workers. *Mgmt. Sci.*, 1962, **8**, 359-370.
- GAGE, N. L., & CRONBACH, L. J. Conceptual and methodological problems in interpersonal perception. *Psychol. Rev.*, 1955, **62**, 411-422.
- GUETZKOW, H., & FOREHAND, G. A. A research strategy for partial knowledge useful in the selection of executives. In R. Tagiuri (Ed.), *Research needs in executive selection*. Boston: Harvard University, Graduate School of Business Administration, 1961.
- GUETZKOW, H., FOREHAND, G. A., & JAMES, B. J. An evaluation of educational influence on executive judgment. *Admin. Sci. Quart.*, 1962, **6**, 484-500.
- HENRY, E. R. Some points of view about research on executive selection. In R. Tagiuri (Ed.), *Research needs in executive selection*. Boston: Harvard University, Graduate School of Business Administration, 1961.
- HUMPHREYS, L. G. A note on the multitrait-multimethod matrix. *Psychol. Bull.*, 1960, **57**, 86-88.
- KAY, B. R. The use of critical incidents in a forced-choice scale. *J. appl. Psychol.*, 1959, **43**, 269-270.
- MARCH, J. G., & SIMON, H. *Organizations*. New York: Wiley, 1958.
- MAYO, G. D. Peer ratings and halo. *Educ. psychol. Measmt.*, 1956, **16**, 317-323.
- NELSON, C. A look at some of the basic organizational forces that affect leadership in management and may cause management training courses to fail. Unpublished manuscript, Management Research Associates, Chesterton, Indiana, 1959.
- TAGIURI, R. Survey of recurrent issues. In R. Tagiuri (Ed.), *Research needs in executive selection*. Boston: Harvard University, Graduate School of Business Administration, 1961.
- TANNENBAUM, R., WESCHLER, I. R., & MASSARICK, F. *Leadership and organization: A behavioral science approach*. New York: McGraw-Hill, 1961.
- THORNDIKE, R. L. Some methodological issues in executive selection. In R. Tagiuri (Ed.), *Research needs in executive selection*. Boston: Harvard University, Graduate School of Business Administration, 1961.

(Received July 19, 1962)

KNOWLEDGE OF RESULTS AND SIGNAL RATE IN MONITORING:

A TRANSFER OF TRAINING APPROACH¹

EARL L. WIENER

University of Miami

This study investigated the transfer effects of training with 3 signal rates and 3 levels of knowledge of results (KR) in a visual monitoring task. Each S monitored for 48 min. under 1 of 9 signal rate-KR conditions on Day 1. On Day 2 all Ss monitored under the medial signal rate with no KR. Results show: (a) on Day 1 mean probability of detection increased with signal rate and amount of KR, (b) these differences persisted on Day 2 when KR was withdrawn, and (c) commissive errors were higher with partial KR than with either full KR or none. It is concluded that training a monitor with KR and high signal rates may improve performance when he must monitor with low signal rates and no feedback.

While considerable research on the effects of signal rate and knowledge of results (KR) has been reported, there has been little or no effort to determine the permanence or transfer of these effects. Numerous authors (Deese & Ormond, 1953; Harabedian, McGrath, & Buckner, 1960; Jenkins, 1958; Kappauf & Powe, 1959; Nicely & Miller, 1957; Pollack & Knaff, 1958a) have shown that an increase in signal rate results in an increase in probability of signal detection. A recent study by Floyd, Griggs, and Baker (1961) transferred subjects from signal rates of 30, 10, and 3 equally spaced signals to a second session of 10 equally spaced signals. The group receiving 10 signals during the first session was clearly superior to the other two on the second session.

Superior performance by groups receiving KR has been demonstrated by Baker (1959, 1960a), Garvey, Taylor, and Newlin (1959), Loeb and Schmidt (1960a, 1960b), Mackworth (1950), McCormick (1959), Pollack

and Knaff (1958b), and Weidenfeller, Baker, and Ware (1962). Mackworth's study (1950) suggested the possibility of a persistent effect of KR when such feedback was later withheld.

In an actual monitored system, KR would be unavailable, or at best, extremely delayed. If the circuitry, human or electronic, to provide immediate KR were present, the occurrence of a signal would have to be known, and thus there would be no need for the human monitor. Therefore, KR is somewhat of an academic question in detection tasks, unless one takes a training point of view. It is entirely feasible to construct a monitoring simulator for training purposes, allowing for feedback of results and manipulation of signal rate. This experiment was designed to determine the transfer effect of various signal rates and levels of KR to a condition of medial signal rate and no feedback.

METHOD

Apparatus

The task was an adaptation of the standard Mackworth clock task (Mackworth, 1944). The display was a Standard Laboratory Timer with 1 rps sweep hand, the instrument face being reversed so that behind the sweep hand was a uniform white surface. Each clock was placed in a wooden box with a Plexiglas window and shock-mounted inside the box with foam rubber. This display is illustrated in Figure 1.

The movement of the sweep hand was controlled by a series of relays and timers, so that a normal

¹ This experiment is based on a dissertation presented to the Graduate School of the Ohio State University. The author is indebted to his joint advisers, G. E. Briggs and Daniel Howland. Primary support for this research was a grant from the Engineering Experiment Station of Ohio State. The basic vigilance apparatus was provided on loan from the Behavioral Sciences Laboratory of the United States Air Force Aeronautical Systems Division, Wright-Patterson Air Force Base. A complete report of this research has been published by that group as AMRL-TDR-62-82.

pulse delivered to the clock advanced the hand $\frac{1}{32}$ of a revolution, or 11.25 degrees, once each second. The signal for which the subject was monitoring was a jump of $\frac{1}{18}$ of a revolution, or 20 degrees. For convenience, this will be referred to as a "double jump," though it is actually only 1.78 times the normal jump. This ratio was determined experimentally in a preliminary study. The customary ratio of 2.00 was found to be too easy a task even without KR.

The occurrence of a double jump was determined by a schedule punched into standard 5-channel teletype tape, stepped through a Western Union Model 24-B tape reader. The apparatus allowed three subjects to be run at once independently: all clocks received signals at the same time.

The KR was presented automatically. The occurrence of a signal started a timer in the KR apparatus. If a response were made during the allotted time (3.5 seconds), the left light remained illuminated (green) as long as the subject held his switch closed. If a response were made when the timer was not timing out the 3.5 seconds, it was a commissive error and the center light illuminated (red) as long as the switch was held. If no response were made to a signal during the allotted time, at the end of the interval the right-hand light illuminated (amber) for 5 seconds. This interval was regulated by a time-delay relay. Information feedback for each subject was independent of that of the other subjects. The KR display was mounted in a green metal box above the clock box, as shown in Figure 1. Each signal and each subject's responses were recorded on an Esterline-Angus 20-pen Model AW recorder.

Each subject sat in a booth at a desk which held the clock and KR display. The subjects responded with a hand-held Packard-Bell SW-141-K silent, squeeze-type switch. The subjects were isolated from each other by the walls of the booth, the three booths being separated from the rest of the room by a black curtain. Auditory isolation was achieved by having subjects wear earphones playing white noise at 75 decibels as measured at the earpiece by a Hermon Hosmer Scott Type 410 sound-level meter. The subjects were not able to detect the occurrence of a signal or a response from other subjects or from apparatus sounds. The master control which pulsed the clocks and the KR apparatus were placed in soundproof Celotex boxes.

Signal Schedules

Three signal schedules were used in this experiment: 16, 32, and 48 signals per 48-minute session. The schedules were made up by determining the intervals between signals with a table of random numbers from a uniform distribution. The following restrictions were imposed on the random assignments: (a) the signals were equally assigned to four 12-minute blocks within the 48-minute session, (b) no intersignal interval was shorter than .3 minute, and (c) no signal occurred during the first minute of a session.

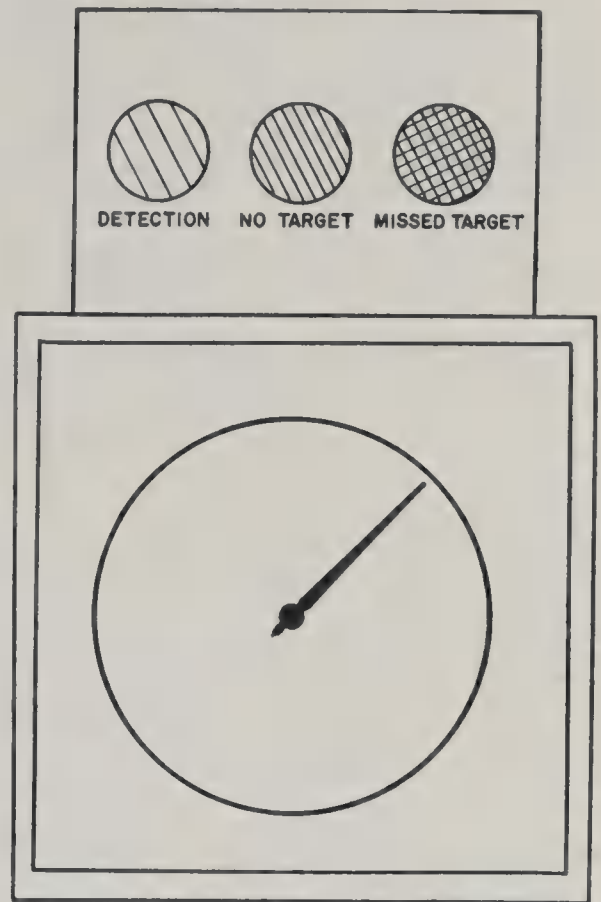


FIG. 1. Mackworth clock and KR display.

Subjects

Subjects were 96 female undergraduate students at Ohio State University. They were recruited for the experiment by various means, including newspaper ads and posters. None had previously served in any type of vigilance experiment. Subjects were paid \$2.00 for their participation. The data from six subjects were not used in the analysis due to apparatus failure (three subjects) and failure of three subjects to carry out their instructions.

Experimental Design

The experimental design for each session was a Lindquist Type III, involving 3 signal rates, 3 KR conditions, and 4 12-minute time blocks, with 10 subjects nested in each of the 9 signal rate-KR combination groups, running under all 4 time blocks, for a total of 90 subjects (Lindquist, 1953).

The three KR conditions are identified by Roman numerals. They are:

- I. No KR
- II. Knowledge furnished only upon the subject's response (i.e., correct detections and commissive errors)
- III. Correct detections, commissive errors, and omissive errors

The full KR condition (III) was included as an ideal which could be attained in a training device. The partial KR condition (II) simulates many real-world systems where immediate or slightly delayed KR can be obtained when a response is made, but knowledge of a missed signal may be impossible to obtain, or at best highly delayed.

On Day 2, all subjects were run under the no-KR condition with 32 signals. This experimental condition is identical to that of Group I-32 on Day 1, except that the actual signal schedule was reversed. For purposes of analysis of the transfer effects, subjects on Day 2 were identified by their Day 1 experimental group, and an identical analysis of variance was performed on the data from Day 2.

Procedure

The subjects were randomly assigned to the nine groups by means of a random number table and were given standard instructions appropriate to their group. Groups run under KR Condition I were told nothing about KR and the display was not present. On Day 2 the KR display was removed and groups run under KR Conditions II and III were simply instructed that the session would be as before, with the exception that no KR would be presented. At no time on either day was signal rate discussed, except that all groups were told that the signals would be very infrequent. Watches were taken from the subjects before each session so that they could not keep time. They knew only that the session would last about an hour.

All subjects were cautioned to stay alert for all double jumps; and at the same time were warned

that "false alarms" would count against them. No attempt was made to set relative weights on the importance of these errors, but only to impress upon the subject the great importance of both. No subject asked about relative weights or penalties for "guessing."

The subjects were given a certain amount of familiarization with the task. In this familiarization, which amounted to about 10 signals appearing in rapid order, subjects sat together in a booth. Both their responses and those of the experimenter furnished some KR. The subjects were warned that the signals occurring during their actual session would be far less frequent than during the instruction period.

RESULTS

Detected Signals

The raw data in terms of percentage signals detected were transformed to radians by the arc-sine transformation (Bartlett, 1947). All conclusions reported will be based on the transformed data. All figures are drawn in terms of percentage detections.

Figures 2 and 3 show the percentage detection as a function of time on Day 1 and Day 2. These separate figures report the same data: in Figure 2 the parameter of the curves is the three KR conditions averaged across the three signal rates. Figure 3 is just the opposite. Figure 4 shows the overall

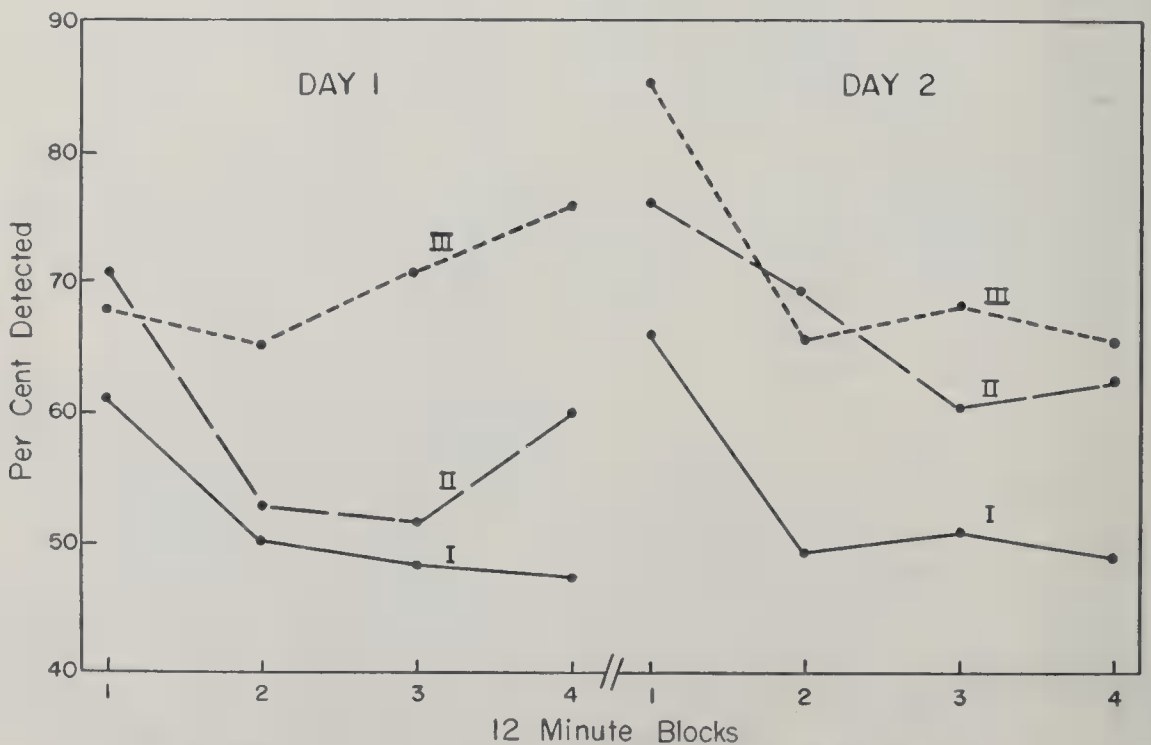


FIG. 2. Percent detection as a function of length of watch by KR groups.

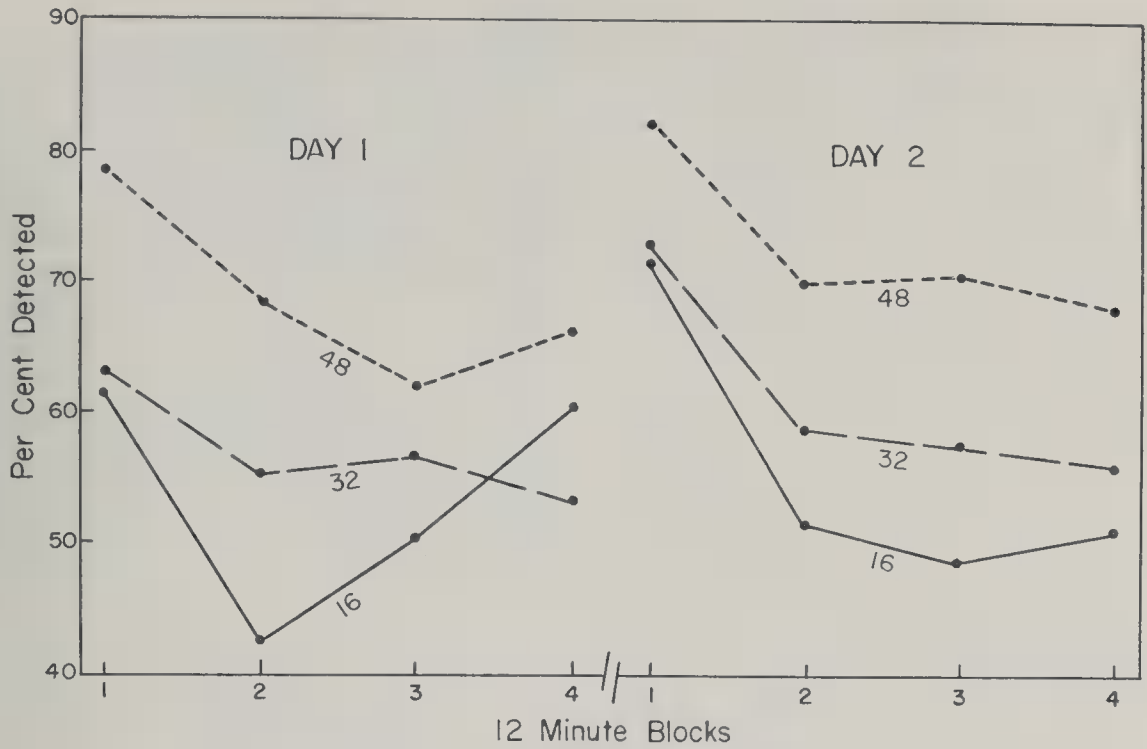


FIG. 3. Percent detection as a function of length of watch by signal rate groups.

detection rate as a function of time for each day, summed over the nine groups. The means of each group and the time block are displayed in Table 1.

The analysis of variance for each day separately is presented in Tables 2 and 3. Table 2 indicates that on Day 1 all main

effects (time, KR, and signal rate) were statistically significant as were both two-way interactions involving time. Table 3 indicates that on Day 2 only the three main effects were significant.

A two-way analysis of variance was performed on the differences between days, and

TABLE 1
MEAN PERCENTAGE OF DETECTED SIGNALS BY SIGNAL RATE, KR CONDITION, TIME BLOCK, AND DAY

Signal rate	Day 1 Time block				Day 2 Time block			
	1	2	3	4	1	2	3	4
KR I								
16	62.5	45.0	52.0	50.0	69.0	35.9	36.2	43.9
32	51.2	42.4	36.1	34.9	57.7	40.0	37.5	37.5
48	77.5	64.1	52.8	56.8	75.2	69.0	62.7	62.7
KR II								
16	70.0	35.0	40.0	65.0	81.5	71.3	61.4	64.0
32	61.3	56.3	56.4	49.9	67.6	67.8	53.8	61.4
48	81.8	66.6	56.7	65.8	82.7	70.1	66.5	66.3
KR III								
16	52.0	47.5	60.0	70.0	67.6	51.2	48.6	49.8
32	76.4	71.5	80.2	79.1	93.9	70.3	78.9	72.7
48	78.2	76.7	76.7	80.1	92.7	73.8	76.4	76.3

TABLE 2
ANALYSIS OF VARIANCE OF PERCENTAGE
OF DETECTIONS ON DAY 1

Source	df	MS	F
Between subjects	89		
Feedback (F)	2	6.54	6.14**
Signal rate (S)	2	4.74	4.44*
F × S	4	1.49	1.40
Error (b)	81	1.07	
Within subjects	270		
Time periods (T)	3	1.77	5.70***
T × F	6	.91	2.94**
T × S	6	.77	2.48*
T × F × S	12	.20	—
Error (w)	243	.31	
Total	359		

* $p < .05$.
** $p < .01$.
*** $p < .001$.

time blocks, as shown in Figure 4. For this analysis, the 10 subjects in each KR-signal rate group were averaged for each time block. Thus nine scores were available for each time period of each day for the analysis presented in Table 4. The time blocks were significantly different, but the differences between days and the day-by-block interaction were not.

To further explore the decremental function, a Duncan multiple range test (Duncan,

TABLE 3
ANALYSIS OF VARIANCE OF PERCENTAGE
OF DETECTIONS ON DAY 2

Source	df	MS	F
Between subjects	89		
Feedback (F)	2	7.35	7.83***
Signal rate (S)	2	6.25	6.65**
F × S	4	2.19	2.33
Error (b)	81	.94	
Within subjects	270		
Time periods (T)	3	4.95	24.65***
T × F	6	.38	1.88
T × S	6	.09	—
T × F × S	12	.13	—
Error (w)	243	.20	
Total	359		

** $p < .01$.
*** $p < .001$.

1955) was performed on the means for each day (Figure 4). On both days, the mean of the first time block was significantly different from all others, and none of the remaining three time block means produced significant differences. This further confirms the view of Jerison and Wallis (1957a, 1957b) that vigilance decrement occurs early in the watch period.

To determine the day-to-day reliability of the subjects' scores, the Pearson product-moment correlation of each subject's detection rate (averaged over the four time periods) between Day 1 and Day 2 was computed. This was found to be $r = .67$, significant at the .001 level for $n = 90$.

TABLE 4
ANALYSIS OF VARIANCE OF PERCENTAGE DETECTED
SIGNALS AS A FUNCTION OF TIME:
DAY 1 AND DAY 2

Source	df	MS	F
Time periods (T)	3	5833	4.59**
Days (D)	1	2415	1.90
D × T	3	880	—
Residual	64	1272	
Total	71		

** $p < .01$.

Commissive Errors

A commissive error was defined as any response occurring more than 3.5 seconds after the previous signal. Thus the subject was virtually unbounded in the number of commissive errors she could make. The commissive error data by groups are shown in Table 5. The raw data revealed an extreme positive skewness. Summing across the four time blocks for each subject, it was found that the majority of subjects made three or fewer commissive errors, but the mean number of such errors on Day 1 was 12.9, due to a vast number made by a few subjects. The percentage of the total 1,158 commissive errors on Day 1 attributed to the eight worse offenders is shown in Figure 5. One can see from this plot that 5 out of 90 subjects (about 6%) accounted for 49% of the commissive errors on Day 1.

An attempt was made to fit a Poisson distribution to the data as a preliminary step toward a possible square root transformation, but this was abandoned after it was shown that the data were by no means Poisson distributed. Again, this was due to the fact that while the mean was large, the scores were clustered near zero. The large value of the mean caused the expected values of the Poisson to be very small in the region near zero, resulting in significant deviations in a chi square goodness-of-fit test.

A preliminary Kruskal-Wallis test on the four time blocks revealed no significant differences due to time. Therefore further

TABLE 5

COMMISSIVE ERRORS BY SIGNAL RATE
AND KR CONDITION

Signal rate	KR Group			Total
	I	II	III	
Day 1				
16	186	104	18	308
32	108	267	41	416
48	95	274	65	434
Total	389	645	124	1158
Day 2				
16	54	39	9	102
32	44	99	31	174
48	58	103	52	213
Total	156	241	92	489

analyses were conducted on the total number of commissive errors made by each subject during the 48-minute run, the 90 subjects being ranked with respect to this measure.

Since no acceptable two-way nonparametric test was available, two separate Kruskal-Wallis tests were performed on the data, and these summarized in Table 6, one to test the signal rate effect and one to test the KR conditions.

Table 5 shows that the number of commissive errors increased with signal rate on both days, but the difference was significant only on Day 2. The effect of KR conditions was significant on both days. This effect was interesting, in that Condition II, partial KR,

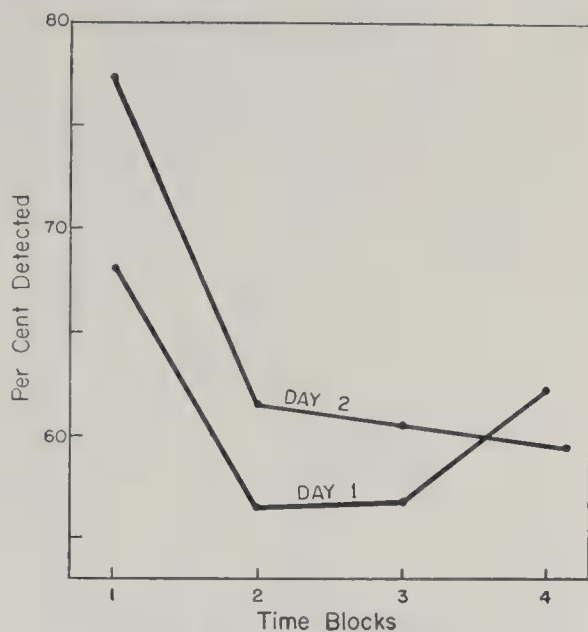


FIG. 4. Percent detection as a function of length of watch, all groups combined.

was accompanied by a large increase in commissive errors; while Condition III, full KR, showed far fewer errors than either of the other conditions.

The Day 1-Day 2 correlation of the subjects' performance was computed by Spearman's nonparametric method. This analysis yielded a rho of .66, significant at the .001 level for $N = 90$.

Two analyses were conducted on the data

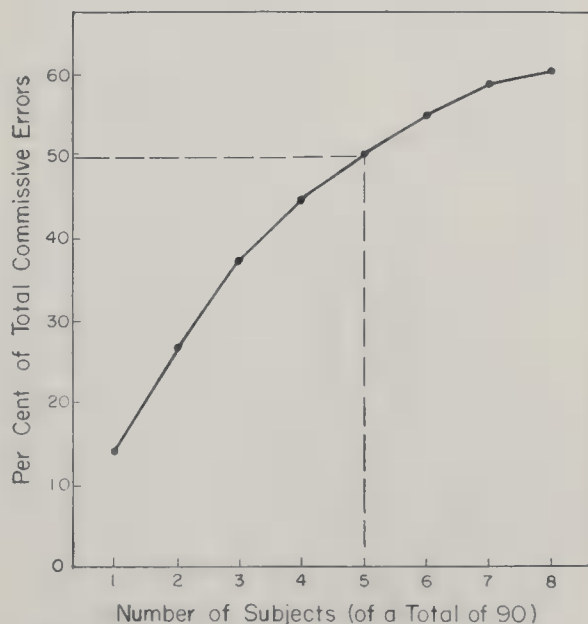


FIG. 5. Percent of the total 1,158 commissive errors on Day 1 contributed by 8 of 90 subjects.

TABLE 6
CHI SQUARE VALUES OF THE KRUSKAL-WALLIS TESTS
FOR COMMISSIVE ERRORS

	Day 1	Day 2	df
Signal rate	1.8	7.0*	2
KR	12.4**	6.6*	2

■ $p < .05$.
** $p < .01$.

from Day 1 merely as a check on the experimental procedures. These were chi square tests of goodness-of-fit to determine whether signal detection was affected by (a) the three stations used by the subjects or (b) the size of the groups that were run. In each case, the expected number of detected signals, based on the number of subjects run at each station or in each size group, was compared to the obtained frequency. Both analyses resulted in nonsignificant chi squares indicating no departure from internal consistency in the experimental procedure.

DISCUSSION

The major results of this experiment support the use of KR as a training aid. The results of Day 1 support the findings of other experimenters that KR favorably affects probability of detection. But more interesting are the results from Day 2, which show that even when KR is withdrawn, the groups initially exposed to this feedback continue to perform at a level which is superior to that of the control group. This divergence between groups, as shown in Figure 2, is encouraging from a training point of view. Whether this difference would persevere through long periods without KR is a question for further research.

It is interesting to note the change in the relative position of groups first exposed to Condition II, partial KR, from Day 1 to Day 2. On Day 1 those groups more closely resembled the control groups than the full KR groups. However, on Day 2, the KR groups tended to coincide and were clearly different from the control. This may indicate that partial KR, which is furnished only when the operator responds, is just as effective as the full KR condition which requires

furnishing information about missed signals.

The signal rate effect further supports the unanimous finding that higher signal rates lead to higher detection rates. The interesting point is again the transfer effect. When all groups were transferred to a 32-signal schedule for Day 2, the groups trained with higher signal rates were clearly superior. Further, the effect of time was uniform over the signal rates on Day 2. The marked Time \times Signal Rate interaction found on Day 1 was absent on Day 2 when all three curves exhibited the usual vigilance decrement. The data presented strongly suggest training monitors with schedules of high signal rate even when they are to be transferred to low signal rate environments.

There are several results which are disappointing from a training point of view. First there is the absence on either day of a Signal Rate \times KR interaction. Since the amount of feedback is dependent on the appearance of signals (with the exception of commissive error feedback), certainly it is reasonable to expect that higher signal rates would enhance the KR effect, but this was not supported by the data.

As the full-KR group improves its performance, detecting more signals, it denies itself the missed-signal feedback which distinguishes it from Condition II. Thus, missed-signal feedback, if effective, is self-eliminating. On this basis one might predict that if KR can produce a steady increase in performance, Conditions II and III will converge in information content and possibly in performance. There is still, however, this difference: subjects in Condition III know that they will be informed if they miss a signal, for whatever this is worth at high performance levels.

Comparison of the 2 days lends further support to previous studies which report little or no performance change from session to session, the most recent being those of Baker, Sipowicz, and Ware (1961) and Ware, Sipowicz, and Baker (1961). Furthermore, the nonsignificance of the Day \times Block interaction demonstrates that the subjects did not acquire a resistance to vigilance decrement with practice. These comparisons substantiate the view that vigilance tasks are fundamen-

tally different than active tasks, and practice alone is not sufficient to elevate performance.

The highly skewed distribution of commissive errors makes generalization very difficult. Indeed, most authors in the field of vigilance have either ignored commissive errors entirely in their reports, or have passed them off with the statement that very few false reports were made.

The curious finding presented here is that partial KR results in a marked increase in commissive errors over the control condition, while full KR results in a reduced number of such errors. Some attempt to explain the large number of commissive errors under Condition II seems worth while. Under the control conditions, the subject received no information regardless of what action or inaction she chose to take. However, under Condition III the subject received full information regardless of her decision. If the subject responded she was able to determine immediately whether that response was right or wrong, and if she did not respond, the missed signal light or its absence gave the same information. But in Condition II, the subject received information only by responding. Thus if the subject were doubtful about whether a presentation was signal or noise, she could resolve this ambiguity by responding. So the large number of commissive errors found under this condition may be accounted for by the subject's desire for information, which could be satisfied only by responding. What is more, this tendency transferred to Day 2 even when no information was available. One might say that partial KR as presented here (as well as in practical situations) encourages false responses. It was pointed out previously that the two KR conditions performed equally with respect to detections following transfer to a no-KR condition. The difference in probability of false response recommends the addition of missed signal information during early training.

Deese's expectancy theory (1955) states, in brief, that the likelihood that the subject will respond to a randomly occurring signal depends on his expectancy about the appearance of a signal. This expectancy is built up as a kind of "averaging process" based on past signals. The Baker (1958, 1960b, 1961)

extension of this theory is based on the notion of the observer perceiving the true temporal nature of the signal schedule.

The expectancy theory does not attack the problem of transfer of training directly, but it is not difficult to make predictions about transfer consistent with the general theory. As for signal rate, one would predict that the higher the signal rate during training, the higher the expectancy during transfer, and thus the better the performance. This is supported by the data. However, there is a slight inconsistency here. The expectancy theory would predict also that the groups transferred to the identical signal rate schedule would have the most accurate perception of the intersignal interval, and thus the theory would predict that the 32-signal groups would perform better during transfer than the other two groups, the effect demonstrated by Floyd, Griggs, and Baker (1961). Under the transfer condition on Day 2, the 48-signal groups were presented with fewer signals than on Day 1, but they carried their expectancy with them into Day 2, accounting for their high performance, even on a different signal schedule. The 16-signal groups were at a disadvantage from either point of view. They suffered from both low expectancy and a change of signal schedule when transferred to the 32-signal condition.

The disagreement between the findings of the present study and that of Floyd, Griggs, and Baker may lie in the fact that their signals were evenly spaced over time, allowing maximal information about intersignal interval, with no need for the "averaging" process discussed by Baker.

Though the expectancy theory does not deal with commissive errors, it again seems consistent with the theory to predict that under transfer, the higher signal rate groups would commit more false reports. One might say that under transfer, a large number of commissive errors is the price of building in a high expectancy. This view is supported by the data, which show a nonsignificant signal rate effect on Day 1 and a significant effect in the predicted direction on Day 2.

Some implications of this research for practical application should be clear. Primarily they consist of: (a) Initially training

monitors with full KR; (b) initially training monitors with high signal rates. Just how high this should be cannot be answered by this study.

The question of permanence need not be the deciding factor in determining feasibility of such training. It is entirely possible to retrain monitors at regular intervals after they have been on the job without KR. In this manner it may be possible to "refresh" the monitor with periodic re-exposure to a monitoring training device with full KR feedback.

REFERENCES

- BAKER, C. H. Vigilance: Two tentative theoretical approaches. Canadian Delegation Memorandum No. CWS (P) /P (58) 18, 1958, Commonwealth Advisory Committee on Defense Science, Toronto.
- BAKER, C. H. Three minor studies of vigilance. *Def. Res. Med. Lab. Rep.*, Toronto, 1959(Apr), No. 234-2.
- BAKER, C. H. Maintaining the level of vigilance by means of artificial signals. *J. appl. Psychol.*, 1960, 44, 336-338. (a)
- BAKER, C. H. Toward a theory of vigilance. In A. Morris & E. P. Horne (Ed.), *Visual search techniques*. Washington: National Research Council, 1960. (b)
- BAKER, C. H. Further toward a theory of vigilance. *Def. Res. Med. Lab. Rep.*, Toronto, 1961, No. 234-10.
- BAKER, R. A., SIPOWICZ, R. R., & WARE, J. R. Effects of practice on visual monitoring. *Percept. mot. Skills*, 1961, 13, 291-294.
- BARTLETT, M. S. The use of transformations. *Biometrics*, 1947, 3, 39-52.
- DEESE, J. Some problems in the theory of vigilance. *Psychol. Rev.*, 1955, 5, 359-368.
- DEESE, J., & ORMOND, E. Studies of detectability during continuous visual search. *USAF WADC tech. Rep.*, 1953, No. 53-8.
- DUNCAN, D. B. Multiple range and multiple *F* tests. *Biometrics*, 1955, 11, 1-42.
- FLOYD, A., GRIGGS, G. D., & BAKER, R. A. Role of expectancy in auditory vigilance. *Percept. mot. skills*, 1961, 13, 131-134.
- GARVEY, W. D., TAYLOR, F. V., & NEWLIN, E. P. The use of "artificial signals" to enhance monitoring performance. *USN Res. Lab. Rep.*, 1959, No. 5269.
- HARABEDIAN, A., MCGRATH, J. J., & BUCKNER, D. The probability of signal detection in a vigilance task as a function of inter-signal interval. Technical Report No. 3, 1960, Human Factors Research, Incorporated, Human Factor Problems in Anti-Submarine Warfare, Los Angeles, California.
- JENKINS, H. M. The effect of signal rate on performance in visual monitoring. *Amer. J. Psychol.*, 1958, 71, 647-661.
- JERISON, H. J., & WALLIS, R. A. Experiments on vigilance: II. One-clock and three-clock monitoring. *USAF WADC tech. Rep.*, 1957, No. 57-206. (a)
- JERISON, H. J., & WALLIS, R. A. Experiments on vigilance: III. Performance on a simple vigilance task in noise and quiet. *USAF WADC tech. Rep.*, 1957, No. 57-318. (b)
- KAPPAUF, W. E., & POWE, W. E. Performance decrement at an audio-visual checking task. *J. exp. Psychol.*, 1959, 57, 49-56.
- LINDQUIST, E. F. *Design and analysis of experiments in psychology and education*. Boston: Houghton Mifflin, 1953.
- LOEB, M., & SCHMIDT, E. A. A comparison of the effects of different kinds of information in maintaining efficiency on an auditory monitoring task. *USA Med. Res. Lab. Rep.*, 1960, No. 453. (a)
- LOEB, M., & SCHMIDT, E. A. Influence of time on task and false information on efficiency of responding to pure tones. *USA Med. Res. Lab. Rep.*, 1960, No. 426. (b)
- MCCORMACK, P. D. Performance in a vigilance task with and without knowledge of results. *Canad. J. Psychol.*, 1959, 13, 68-71.
- MACKWORTH, N. H. Notes on the clock test: A new approach to the study of prolonged visual perception to find the optimum length of watch for radar operators. *Med. Res. Council, Appl. Psychol. Unit Rep.*, 1944, No. 1.
- MACKWORTH, N. H. Researches on the measurement of human performance. *Med. Res. Council, spec. Rep. Ser.*, 1950, No. 268.
- NICELY, P. E., & MILLER, G. A. Some effects of unequal spatial distribution on detectability of radar targets. *J. exp. Psychol.*, 1957, 53, 195-198.
- POLLACK, I., & KNAFF, P. R. Effect of rate of target presentation on target detection probability. *Amer. Psychologist*, 1958, 13, 414. (Abstract) (a)
- POLLACK, I., & KNAFF, P. R. Maintenance of alertness by a loud auditory signal. *J. Acoust. Soc. Amer.*, 1958, 30, 1013-1016. (b)
- WARE, J. R., SIPOWICZ, R. R., & BAKER, R. A. Auditory vigilance in repeated sessions. *Percept. mot. Skills*, 1961, 13, 127-129.
- WEIDENFELLER, E. W., BAKER, R. A., & WARE, J. R. Effects of knowledge of results (true and false) on vigilance performance. *Percept. mot. Skills*, 1962, 14, 211-215.

(Received July 23, 1962)

A NOTE ON JOB PERFORMANCE: DIFFERENCES BETWEEN RESPONDENT AND NONRESPONDENT SALESMEN TO AN ATTITUDE SURVEY

WAYNE K. KIRCHNER AND NANCY B. MOUSLEY

Minnesota Mining and Manufacturing Company, St. Paul

Salesmen respondents ($N=72$) and nonrespondents ($N=19$) to a mail attitude questionnaire were compared in terms of 2 objective measures of performance: net sales points and net total points. Mean scores on both measures were significantly higher for respondents than for nonrespondents. These results tended to follow results of other studies in nonindustrial settings that suggested volunteers or respondents are, in general, "better" persons in terms of such variables as motivation, personality and, in this case, job performance.

Even a cursory search of the literature would reveal that comparisons of volunteers versus nonvolunteers or respondents versus nonrespondents in experiments and surveys of all kinds are fascinating topics for psychologists. The number of such studies is immense. For example, Lubin, Levitt, and Zuckerman (1962) have shown that nurses who reply to a survey questionnaire differ in personality characteristics (i.e., more orderly, more dependent) from nonrespondents. Bair and Gallagher (1960) have shown that volunteers for hazardous duty in the United States Navy differ from nonvolunteers in terms of being more likely to complete flight training and excelling more in leadership. Burchinal (1960) found differences in personality among students who volunteered to complete a questionnaire at night versus those who completed it during a regular class session.

In general, these studies and others by Rosen (1951), Donald (1960), Blake, Berkowitz, Bellamy, and Mouton (1956) and Siegman (1956)—to cite only a few—show that there are definite, although not necessarily consistent, differences between volunteers and nonvolunteers or respondents and nonrespondents in terms of personality, demographic data, and motivation. Subjectively, it appears that in most studies, volunteers or respondents are "good guys" and nonvolunteers or nonrespondents, "bad guys."

In any case, none of these studies have been aimed particularly at comparing re-

spondents versus nonrespondents in terms of actual job performance nor have they been conducted in industrial settings. It would also seem of value to see whether differences exist between respondents and nonrespondents in terms of how they actually perform on the job.

Specifically, this brief note summarizes such a comparison. This was a comparison of performance on sales jobs by respondents and nonrespondents to any attitude survey.

METHOD

Early in 1962, attitude questionnaires were mailed to 91 salesmen in one division of a large, midwestern manufacturing concern. These were typical attitude measures encompassing items related to such things as training, benefits, general work attitudes, etc. This survey was conducted strictly on a volunteer basis and after two follow ups, 72 salesmen (79.1%) responded, leaving 19 nonrespondents. By means of a coding method, it was possible to determine which salesmen had not replied.

The two groups of respondents and nonrespondents were then compared in terms of two objective measures of performance: (a) net sales points; (b) net total points. Net sales points were gained by direct selling of the product; net total points included net sales points plus points gathered in other ways (collection of delinquent accounts, etc.).

RESULTS AND DISCUSSION

Results are shown in Table 1. As is seen, respondents do differ significantly from nonrespondents in terms of both net sales points and net total points achieved.

For example, respondents average 123.5 more net sales points higher and 102.8 more

TABLE 1
COMPARISON OF RESPONDENTS AND NONRESPONDENTS TO ATTITUDE SURVEY ON
SALES PRODUCTION AND TOTAL PRODUCTION

	Net sales points			Net total points		
	Number ^a	<i>M</i>	<i>SD</i>	Number	<i>M</i>	<i>SD</i>
Respondents	68	307.4	235.1	72	533.2	201.8
Nonrespondents	17	183.9	122.4	19	430.4	163.3
	<i>t</i> = 3.00**			<i>t</i> = 2.32*		

^a The number of salesmen having net sales points is slightly smaller than the total *N* because some salesmen did not engage in any selling activities in this period.
* *p* < .05.
** *p* < .01.

net total points higher than nonrespondents. The respondents then tend to be the good guys or the better salesmen.

This, of course, suggests that the attitude results may be biased. It is possible that respondents give more favorable responses because they are better performers and presumably more satisfied.

In any case, respondents do differ in this study from nonrespondents in terms of actual job performance with the findings reinforcing the oft-found result from past research that respondents are generally "better." This now seems true as well in an actual work situation in an industrial setting.

REFERENCES

BAIR, J. R., & GALLAGHER, T. J. Volunteering for extrahazardous duty. *J. appl. Psychol.*, 1960, **44**, 329-31.

BLAKE, R. R., BERKOWITZ, H., BELLAMY, R. Q., & MOUTON, JANE S. Volunteering as an avoidance act. *J. abnorm. soc. Psychol.*, 1956, **53**, 154-156.
BURCHINAL, L. G. Personality characteristics and sample bias. *J. appl. Psychol.*, 1960, **44**, 172-174.
DONALD, M. N. Implications of non-response for the interpretation of mail questionnaire data. *Publ. Opin. Quart.*, 1960, **24**, 99-114.
LUBIN, B., LEVITT, E. E., & ZUCKERMAN, M. Some personality differences between responders and nonresponders to a survey questionnaire. *J. consult. Psychol.*, 1962, **26**, 192.
ROSEN, E. Difference between volunteers and non-volunteers for psychological studies. *J. appl. Psychol.*, 1951, **35**, 185-193.
SIEGMAN, A. Responses to a personality questionnaire by volunteers and nonvolunteers to a Kinsey interview. *J. abnorm. soc. Psychol.*, 1956, **52**, 280-281.

(Received July 24, 1962)

Journal of Applied Psychology

KENNETH E. CLARK, Editor
University of Rochester

Table of Contents

Personality Measurement, Faking, and Detection: An Assessment Method for Use in Personnel Selection: Warren T. Norman.....	225
Self-Percept and Consumer Attitudes toward Small Cars: Eugene Jacobson and Jerome Kossoff.....	242
Underlying Sources of Job Satisfaction: Frank Friedlander.....	246
A Note on <i>The Criterion</i> : Marvin D. Dunnette.....	251
The Relationship between Social Desirability and Internal Consistency of Personality Scales: Allen L. Edwards, James A. Walsh, and Carol J. Diers.....	255
Effect of Variation in Task Complexity and Displayed Information on Operator Performance: Kenneth Ziedman and John Lyman.....	260
Job Attitudes in Management: III. Perceived Deficiencies in Need Fulfillment as a Function of Line versus Staff Type of Job: Lyman W. Porter.....	267
A Verification Scale for the Minnesota Vocational Interest Inventory: David P. Campbell and Rachel W. Trockman.....	276
Time, Awareness, and Order of Presentation in Opinion Change: Duane P. Schultz.....	280
Relationships between Carbon Chain Length and Avoidance Responses in Rats: Francis J. Blaisdell.....	284

American Psychological Association

Consulting Editors

GEORGE E. BRIGGS, *Ohio State University*
NORMAN FREDERIKSEN, *Educational Testing Service*
LEONARD D. GOODSTEIN, *University of Iowa*
EDWIN R. HENRY, *Peace Corps*
JOHN HOLLAND, *American College Testing Program*
CLIFFORD E. JURGENSEN, *Minneapolis Gas Company*

LAURENCE S. MCGAUGHRAN, *University of Houston*
QUINN MCNEMAR, *Stanford University*
HAROLD F. ROTHE, *Beloit Corporation*
THOMAS A. RYAN, *Cornell University*
ALEXANDER G. WESMAN, *Psychological Corporation*
CLARK L. WILSON, *Batten, Barton, Durstine, and Osborn, Incorporated*

This journal gives primary consideration to original investigations in any field of applied psychology except clinical psychology, although a descriptive or theoretical article may be accepted if it represents a special contribution in an applied field. Quantitative investigations of interest or value to psychologists working in the following broad fields will be considered: vocational and educational prognosis, diagnosis, and guidance at the secondary and college level; personnel research in business, industry, and government; engineering psychology; industrial working conditions; research on opinion and morale factors; job analysis and classification research; market and advertising research.

Manuscripts must be accompanied by an abstract of 100–120 words typed on a separate sheet of paper. The abstract should conform to the style of *Psychological Abstracts*. Detailed instructions for preparation of the abstracts appeared in the *American Psychologist* (1961, 16, 833), or they may be obtained from the Editor or from the APA Central Office.

Manuscripts should be addressed to the Editor:

Dr. Kenneth E. Clark
College of Arts and Science
302 Morey Hall
University of Rochester
Rochester 20, New York

All manuscripts must be submitted in duplicate. Original figures are prepared for publication; duplicate figures may be photographs or pencil-drawn copies.

Manuscripts must conform to the style requirements described in the *Publication Manual of the American Psychological Association*.

The following policies govern reprinting of materials copyrighted in APA journals: (a) to require approval to reprint for tables and figures, and for text only if more than 500 words [total from one article] in length; (b) to continue approval to reprint, contingent on the author's approval, for articles reprinted in whole or in major part; (c) to negotiate where possible the dedication of royalties from commercial publishers to the American Psychological Foundation.

Journal of Applied Psychology

Published bimonthly by the
American Psychological Association
Prince and Lemon Sts., Lancaster, Pa.
and 1333 Sixteenth Street N. W.
Washington 6, D. C.

\$10.00 per volume

\$2.00 per issue

HELEN ORR
Managing Editor

ELIZABETH S. REED
Advertising Manager

VIRGINIA RICHARDS
Editorial Assistant

Subscriptions, orders, and business communications should be addressed to the American Psychological Association, 1333 Sixteenth St. N.W., Washington 6, D. C. Address changes must reach the subscription office by the tenth of the month to take effect the following month. Undelivered copies resulting from address changes will not be replaced; subscribers should notify the post office that they will guarantee second-class forwarding postage. Other claims for undelivered copies must be made within four months of publication.

Second class postage paid at Lancaster, Pennsylvania and at additional mailing places.

© 1963 by the American Psychological Association, Inc.

Journal of Applied Psychology

VOL. 47, NO. 4

AUGUST 1963

PERSONALITY MEASUREMENT, FAKING, AND DETECTION:

AN ASSESSMENT METHOD FOR USE IN PERSONNEL SELECTION¹

WARREN T. NORMAN

University of Michigan

Usual methods for personality assessment have been found unsuitable for use in personnel selection contexts. An alternative method of item construction and of scoring key and detection scale development for personality inventories is proposed. Results of a double cross-validation study based on 456 male Ss using 3 newly developed forced-choice inventories indicate that (a) over 90% of the test performances can be correctly identified as self-report or faked, (b) mean score profiles under the 2 conditions for the 5 personality variables under study are virtually congruent and variances under the faking condition are uniformly smaller, (c) validities against peer-nomination criteria are in the moderate range for all 5 variables and (d) Kuder-Richardson Formula 20 reliabilities between .83 and .92 were obtained.

Despite the plethora of devices and procedures currently available in the area of personality assessment, few, if any, appear to be suitable for use in personnel selection contexts. Without even raising the critical questions of demonstrable relevance and measurement precision, many have to be excluded on a simple cost basis. Elaborate situational performance tests, highly instrumented (so-called "objective") laboratory techniques, and the vast majority of projective and intensive diagnostic interviewing procedures are simply too expensive even to be considered for routine use in most personnel screening and selection settings.²

¹ Portions of this paper were presented at the meetings of the Fourteenth International Congress of Applied Psychology, 13-19 August 1961 in Copenhagen, Denmark, and at the meetings of the American Psychological Association, 3 September 1962. Material included in this report stems in part from a project sponsored by the Personnel Laboratory, Aeronautical Systems Division, Air Force Systems Command, Lackland Air Force Base, Texas.

² Where questions of cost have been waived temporarily for the sake of research, (e.g., Kelly & Fiske, 1951; and the studies reviewed by Cronbach, 1956, and by Loevinger, 1959), the data obtained hardly warrant any prolonged grief by the personnel psychologist deprived of these techniques.

Self-report personality questionnaires, inventories, and check lists, while not usually precluded by considerations of cost alone, nonetheless do present problems of their own. As Dunnette, McCartney, Carlson, and Kirchner (1962) have recently pointed out, research evidence has been accumulating for a number of years which indicates that for even the most carefully constructed devices of this sort—i.e., those using the most sensitive formats for stimulus presentation, and the most sophisticated methods of scoring key development so far suggested—the effects of attempted faking and "slanting" of responses are pronounced. A detailed analysis of this problem and a review of some of the more pertinent arguments and research findings which may eventually bear on its resolution have been presented elsewhere (Norman, 1961b). To put it briefly, there is little evidence to date to support any of the following as a *sufficient* means of solving or circumventing the problem of faking or response dissimulation on personality inventories and questionnaires: forced-choice sets of stimuli matched either on endorsement frequency or on rated "desirability"; auxiliary correction

or suppressor scales; pseudoability or maximal performance tasks; empirical (i.e., contrasted groups) key construction methods, whether to content-relevant or irrelevant stimuli; or response set and stylistic response tendency measures.

This is not to say that these techniques are generally useless nor that there is no evidence or reasonable argument to support one or more of them in preference to alternative methods for certain purposes. Each of them does in fact represent an interesting and, in some cases, an ingenious effort to deal with the problem of faking or some related issue. The position is taken here, however, that the best that can be said at the present time for any of these approaches taken singly or for any combination of them thus far suggested is that the degree and generality of their effectiveness in dealing with the problems of response contamination and faking in the area of personality assessment is yet to be demonstrated. For some there are sufficient data and theoretical arguments already available to warrant a somewhat stronger statement.

One way to circumvent (rather than to solve) the problem of faking or slanting of inventory responses is well known and has been widely employed. By the judicious use of rating procedures with knowledgeable but disinterested informants, useful data on the personality attributes of the ratees can often be obtained. But for the psychologist concerned with personnel selection, this suggestion presents as many problems as it avoids. For example, although the problem of faked responses may not be so salient in this case as when the subject is also the respondent, still there are probably not many instances in which informants who are sufficiently "knowledgeable" are also "disinterested." And what about the broad class of other response sets (e.g., centrality versus extremity tendencies, inclusiveness versus exclusiveness, etc.), well known for their confounding effects when not all subjects are rated by the same informant or when multiple informants assess, but disagree on, the status of a given subject? The list of problems could be lengthened considerably but there is no need; the re-

views by Cronbach (1946, 1950) and by Guilford (1954, Ch. 11) summarize nicely most of the critical problems of this sort. And of even greater importance, they offer numerous suggestions, based on an extensive body of research findings, by which the effects of many of these contaminants can be eliminated or effectively minimized.³

But even assuming that the major sources of contamination in ratings can, in principle, be effectively suppressed to yield valid and reliable assessments, it is likely that the costs and procedural complexities required to obtain such data would make this approach unfeasible also in the great majority of personnel screening and selection situations.

The research presented below grew out of just such a situation. In the attempt to provide a means for measuring a previously established set of relevant personality attributes, derived from ratings, within the scope of feasibility for a particular selection setting, a method of scoring key construction for self-report inventories was discovered. It is suggested that this method may have considerable generality and utility for other selection contexts because of the way in which it handles the problem of intentional faking or distortion of self-report inventory responses.

CRITERION VARIABLES

In a series of peer nomination rating studies using the scales from Cattell's condensed personality sphere set with groups of normal male subjects, Tupes and Christal (1958, 1961) have demonstrated the existence and relative orthogonality of five personality factors. These findings have been replicated on other samples using modified data collection and analysis methods by Wherry, Stander, and Hopkins (1959) and by Norman (1963). The populations on which these results are based range from United States Air Force officer candidates through college

³ It is interesting to observe for how long a time and to what extent attention has been focused on the problems of response contamination in the area of rating procedures and how well formulated the remedial or prophylactic principles are in that domain compared to the recency of comparable research on self-report inventories and the relative chaos that is so apparent in this area at present.

TABLE 1
ABBREVIATED DESCRIPTIONS OF THE PEER NOMINATION CRITERION RATING
SCALES AND THEIR FACTOR DESIGNATIONS

Factor name ^a	Abbreviated scale labels ^b		
	Number	Pole A	Pole B
I. Extroversion or Surgency	1	Talkative-Silent	
	2	Frank, Open-Secretive	
	3	Adventurous-Cautious	
	4	Sociable-Reclusive	
II. Agreeableness	5	Goodnatured-Irritable	
	6	Not Jealous-Jealous	
	7	Mild, Gentle-Headstrong	
	8	Cooperative-Negativistic	
III. Conscientiousness	9	Fussy, Tidy-Careless	
	10	Responsible-Undependable	
	11	Scrupulous-Unscrupulous	
	12	Persevering-Quitting, Fickle	
IV. Emotional Stability	13	Poised-Nervous, Tense	
	14	Calm-Anxious	
	15	Composed-Excitable	
	16	Not Hypochondriacal-Hypochondriacal	
V. Culture	17	Artistically Sensitive-Artistically Insensitive	
	18	Intellectual-Unreflective, Narrow	
	19	Polished, Refined-Crude, Boorish	
	20	Imaginative-Simple, Direct	

Note.—Based on original findings by Tupes and Christal (1958).
^a Pole A.
^b For the actual scale labels employed in data collection, see Cattell (1947).

undergraduates to applicants for graduate school training programs in clinical psychology. In addition, Tupes (1957) has demonstrated predictive validities for these factors against military officer effectiveness criteria. Table 1 lists the names that have been given to these factors and abbreviated labels for the 20 scales that have been found to load most highly on these factors in prior studies.

The remarkable stability of this five-factor structure, the implications this provides for the development of an adequate taxonomy for personality and results of the use of these scales as predictors for a limited number of other performance criteria have been reviewed elsewhere (Norman, 1963). For purposes of the present discussion, these rating factors are taken as a set of intermediate criteria for use in the development of self-report inventory

measures of these variables and against which such measures are to be “validated.”

PREDICTOR INSTRUMENTS AND THE INITIAL RATIONALE FOR KEY CONSTRUCTION

Despite the reservations listed earlier relative to the use of self-report inventory methods for personality assessment in selection contexts, the available research data, especially those based on forced-choice response formats, did seem to indicate that this approach held some promise of success in this effort. The early work by Jurgensen (1944) had employed endorsement rate (or “preference index”) as a basis for matching stimuli to construct forced-choice items. The later findings by Berkshire (1958) clearly indicated, however, that this had been an unfortunate choice. Thus the susceptibility to faking of the Jurgensen Classification In-

ventory reported in the interim by Longstaff and Jurgensen (1953) and the later but unfortunately similar experience of Maher (1959) could plausibly both be laid to this cause, at least in part.

Berkshire's results did indicate, however, that somewhat more effective matching could be achieved if rated "job importance" of the stimuli were used instead as a matching index. In addition, Edwards (1954, 1957) had also argued in favor of a rated stimulus characteristic (in this case "social desirability") as a basis for constructing well-matched, forced-choice items. However, a review of the literature indicated that the use of *average* importance or desirability ratings as a basis for matching stimuli was probably not a fully adequate criterion for several reasons.

First, it is seldom possible to find stimulus pairs with exactly the same mean ratings. Edwards, who acknowledged this difficulty and settled for a .5 scale unit mean difference (on a 9-point scale) to pair A and B stems for his Personal Preference Schedule items, subsequently found a correlation of about .40 between endorsement frequencies for the A response and the A-B rated social desirability difference values over all items in his test. A more stringent mean "equivalence" criterion could, of course, always be required and, on the basis of the above result, certainly would seem to be warranted. But it still seemed doubtful that other findings of the fakability of the EPPS such as those by Borislow (1958), by Corah, Feldman, Cohen, Gruen, Meadow, and Ringwall (1958), or even those by Edwards, Wright, and Lunneborg (1959) could be attributed wholly to just the looseness of the mean desirability criterion used in constructing this inventory.

A second point is therefore worth noting, i.e., that the use of *mean* ratings as a sole criterion for matching stimuli ignores other possibly relevant parameters of the rating distributions. Unless the variances of the ratings are both near zero (a criterion suggested by Maher, 1959, but one seldom attainable with such data), it is possible that the variances could be appreciably different or, even if not, that the correlation between the ratings of the two stimuli could be quite

low. If either of these situations should obtain, then clearly for at least some of the respondents the two stimuli would be different as regards their importance or desirability parameters for them. The findings by Rosen (1956), Borislow (1958), and Messick (1960) would in fact all seem to indicate that the general social desirability stereotype may well be the composite of a rather disparate set of "personal desirability" stereotypes held by different members of a population. The high correlations obtained by Edwards *between average* desirability scale values from pairs of samples from different populations in no sense precludes the possibility of such *intrasample* heterogeneity of opinion. However, this problem could, in principle at least, also be surmounted by requiring either nearly zero variances or, alternatively, nearly equal variances plus a high positive correlation in order for two stimuli to be paired.

Still a third aspect of this approach is somewhat more troublesome to dispose of, however. I refer here to the point that all of the criteria for matching so far discussed are based on parameters of the distributions of ratings from single-stimulus presentations whereas the final test format requires the respondent to make forced-choice discriminations between pairs of stimuli. This constitutes a major problem for two reasons. First, it is well known that much finer discriminations can be made between stimuli when they are presented together in pair comparisons than when they are presented one at a time. Thus, two items which are indistinguishable in desirability when presented and rated singly may be quite clearly differentiable in desirability when placed together in a pair comparison. Secondly, the use of single-stimulus rating data as a basis for constructing forced-choice items assumes that the desirability parameters of individual stimuli are not affected by being placed in different contexts. This is tantamount to saying, in effect, that it makes no difference what the content of the stimulus is with which a given other stimulus is paired as long as they both have the same single-stimulus desirability parameters. On the face of it this seems improbable. There exists, in fact,

a considerable body of research data in the areas of sensory psychophysics and attitude scaling indicating that perceived attributes of stimuli are markedly affected by variations in the context or background in which they are presented. Unfortunately, these findings are neither sufficiently specific nor extensive enough to permit one to predict the direction and magnitude of the mutual influence any pair of personality descriptive items will have on each other when placed together in a forced-choice format.

Thus it would appear that if one were bent on constructing forced-choice items that were highly resistant to dissimulation tendencies, the most defensible approach on the basis of our present knowledge would be to construct a large number of such items but to retain for scoring only those which remain indistinguishable to respondents in terms of desirability, after they have been put in forced-choice form. By using single-stimulus rating parameters to pair the stimuli it would seem plausible that one could probably maximize the number of well-matched binary items eventually obtained. But one would need to subject all binary items initially formed to a final scaling analysis in order to identify those particular items that could confidently be used for scoring key construction. Only those items (*a*) for which the percentage of endorsement for a given response alternative under both straight, self-report conditions and under instructions to fake or dissimulate were constant (and near 50% as well?) and (*b*) for which the percentage of respondents changing their responses under the two conditions were minimal could be considered well controlled for the contaminant according to this rationale.

On the basis of the considerations and findings outlined above, three forced-choice self-report inventories were developed. For the first of these an initial pool of 342 trait descriptive adjectives was compiled and each was then sorted into one or another of the peer rating criterion factor categories independently by each of three judges. The 193 items for which there was high interjudge agreement as to factor-pole relevance were compiled into booklet form for purposes of

collecting desirability ratings for each term. Each adjective was then rated by a sample of male college students as to the degree of desirability that each felt would be implied by a self-endorsement to that characteristic by a person wishing to be admitted to the United States Air Force Officer Candidate School (OCS) program.⁴ Means and variances of the desirability ratings for each adjective were then computed. Binary forced-choice items were formed by matching terms representing two different factors (based on staff judgments of their content) as closely as possible on these two distribution parameters. In all, 200 binary items were formed; some adjectives being used in more than one forced-choice item. These items were then arranged in a roughly systematic, alternating fashion, typed in booklet form, and the resulting instrument was named the Descriptive Adjective Inventory (DAI).

In the development of DAI, no effort was made to obtain or use information concerning the correlations between ratings of the separate adjectives in forming the forced-choice binary items. However, in constructing two other inventory forms which employ self-report statements instead of single word predicates, this criterion was also utilized to the degree permitted by the data. These two inventories, called the Forced-Choice Self-Report Inventory, Forms A and B (or FCSRI-A and FCSRI-B) contain 192 and 199 binary forced-choice items, respectively. No individual statement appears more than once in these two instruments. A more complete description of the development of these devices has been presented elsewhere (Norman, 1961a). Sample items from DAI and FCSRI forms are shown in Table 2.

In constructing items and initial scoring keys for DAI and FCSRI forms, we relied for item "validities" on our judgments of content relevance of the stimuli. These *a priori* keys were later replaced on DAI by preliminary empirical keys based on a small sample of men from whom both criterion ratings and test performances under straight self-report instructions had been obtained.

⁴ This particular stereotype was chosen for our purposes because of our current research support.

TABLE 2
SAMPLE ITEMS FROM THE DAI AND THE FCSRI-A AND FCSRI-B

DAI	FCSRI
A. Tactful	A. People consider me somewhat of an intellectual.
B. Thorough	B. I am delightful and pleasing to others.
A. Nervous	A. I like my friends to feel sorry for me when I am sick.
B. Preoccupied	B. I feel like a stranger with people.

The principal index used for selecting item categories for each of these preliminary empirical keys was a simple percentage difference in response between median split “highs” and “lows” on each criterion variable. Another sample had taken DAI under both self-report and attempted faking instructions and the percentage of endorsement or “popularity” indices for the items under these two conditions were also referred to when selecting items for these preliminary empirical scales.

The endeavor was to choose only those items for keying which had sizable discrimination indices against the rating criteria and minimal shifts in popularity under the two response conditions. We should have liked also to have had evidence that the validities of the items were maintained under the faked condition. But since criterion data were not available on the latter sample, and fake-take test protocols had not been collected from the former, no analysis pertinent to this question was possible.

VALIDATION OF THE PRELIMINARY EMPIRICAL
KEYS FOR DAI AND WHY THE ABOVE
METHOD OF KEY CONSTRUCTION
IS UNFEASIBLE

The major validation study was carried out at the University of Michigan during the academic year, 1960–61. During the fall semester, 215 men were recruited from fraternity houses on campus and were administered the peer nomination criterion rating scales, DAI, and the two forms of the FCSRI together with a large number of other devices. The total testing time per subject was approximately 15 hours. The DAI and FCSRI forms (plus two other instruments not specifically dealt with in this re-

port) were given twice—first under straight self-report instructions to be as “honest and accurate” in describing oneself as possible, and about 2 weeks later under instructions to respond in the “most desirable manner possible to gain admission to United States Air Force OCS.” Arrangements were made to pay each subject upon completion of the battery at the rate of \$1.00 per hour to insure his cooperation and participation throughout all phases of the testing program.

Criterion rating groups of 6–16 men each were formed within each fraternity in such a way as to maximize interrater familiarity and minimize status differences owing to academic class standing. The DAI and FCSRI forms were filled out individually by each subject at his convenience according to each set of instructions during a 2-week period immediately following receipt of the forms.

During the spring semester, 241 men living in university residence halls completed the same test battery under the same arrangements, bringing the total number of participants in the study of 456.

The ratings and test performances from the 215 fraternity men in the fall semester sample were scored and correlations between the criterion ratings and all preliminary keys for tests in the battery were computed. The results of this analysis for the preliminary empirical keys of DAI (described in the last section) were particularly enlightening. The validities against the criterion ratings for these keys under the straight-take condition were all statistically significant and moderate to low in magnitude (i.e., .40, .25, .40, .24, and .16 for Criterion Variables I-V, respectively). However, the validities for these same keys dropped uniformly to near zero under the faking condition. What is more, the cor-

relations between the straight and fake administration scores on these keys also were nearly equal to zero. Finally, the mean profiles under the two conditions were not congruent (nor even parallel) as is clearly indicated in Figure 1. In this figure the mean profile for the five keys under the straight-take condition is plotted as a horizontal line and the points on the mean profile for the fake-take condition are plotted as deviations in multiples of the standard errors of the straight-take means.

From Figure 1 and the criterion validities described above, it is apparent that for these preliminary keys, one does not have either control of distortions due to faking nor validities which are maintained under both conditions. In fact, the rank order of means under the fake condition is precisely what it should be to gain entrance to OCS if one takes the findings by Tupes (1957) at face value! The fact that the deviations in Figure 1 are negative for Factors V, II, and IV is probably attributable wholly to the partially ipsatized character of these keys and the strong preference for Factor III responses under the faking condition.

Additional light is shed on the problem and a potentially satisfactory solution is sug-

gested, however, if one examines another set of findings from these data.

Upon completion of the testing of the spring semester residence hall sample the 51 rating subgroups from both semesters were divided into a pair of double cross-validation samples (hereafter referred to as A and B) for use in all final test development and data analysis. The groups were assigned to the two samples so as to meet the following criteria as closely as possible. (a) There should be equal numbers of fraternity and residence hall subjects in each sample. (b) There should be equal numbers of freshmen, sophomores, juniors, and seniors or graduate students in each sample. (c) The criterion rating factor structures in each sample should be the same. The extent to which Criteria *a* and *b* were met is indicated in Table 3.

It can be seen from the marginal totals of Table 3 that the two samples are extremely well-matched on class standing and on type of residence. In addition the total numbers for the two samples are exactly equal. Even the differences in frequency within the class by type of residence cells are generally small. The data relevant to Criterion *c* are discussed elsewhere (Norman, 1963). Suffice it to say here that the factor structures for these two

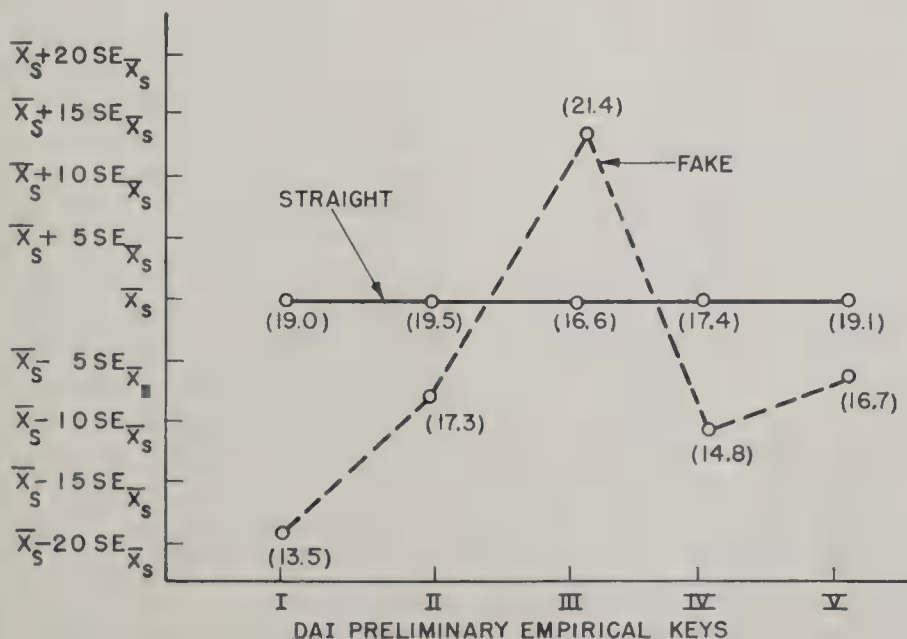


FIG. 1. Comparisons between the mean profiles of scores on the preliminary empirical keys of the DAI under fake-take and straight-take conditions. (Based on 215 fraternity men in the sample. Values in parentheses are means.)

TABLE 3
COMPOSITION OF THE DOUBLE CROSS-VALIDATION SAMPLES A AND B

	Freshmen		Sophomores		Juniors		Seniors & Graduates		Totals	
	A	B	A	B	A	B	A	B	A	B
Fraternities	3	0	51	42	42	45	11	21	107	108
Residence halls	67	70	29	34	13	10	12	6	121	120
Total	70	70	80	76	55	55	23	27	228	228

samples were extremely well matched and that Criterion *c* was met handily.

Responses to all items on DAI, FCSRI-A, and FCSRI-B by each subject under both straight-take and fake-take conditions were punched into IBM cards for item analysis. Each of the double cross-validation samples was then divided into trichotomous criterion classes on each criterion variable with approximately 30% of the sample in the extreme classes and the remaining cases in the middle group. The overall popularity indices and the five matrices of percentage-difference discrimination indices among the trichotomous groups on each criterion variable were then computed separately for each sample for each of the two testing conditions for each test.

A preliminary examination of these data for DAI indicated that for the 200 items in this test: (*a*) the discrimination indices for the straight and fake conditions on each factor in each of the cross-validation samples were correlated approximately zero; (*b*) the correlations between the popularity indices under straight and fake conditions were .24 and .23 in Samples A and B, respectively; (*c*) the correlations between the discrimination indices under the *fake* condition between Sample A and Sample B were essentially zero for all five factors; *but* that (*d*) the correlations between the discrimination indices under the *straight* (self-report) condition between the two samples were moderately high; furthermore, (*e*) the correlations of the popularity indices between the two samples were .97 for the straight condition and .98 for the fake condition!

Because of Findings *a* and *b* it was apparent that not enough (if any) items could

be found which would permit the construction of keys whose validities would be maintained irrespective of the set assumed by the respondent. Findings *c* and *d* implied that the source of the difficulty lay in the fact that the large number of changes in response made between the two conditions were unrelated to criterion factor status. In addition, note that the straight by fake discrimination index correlations—Finding *a*—were zero and *not* negative. Thus, it was not likely that configural response patterns would be any more fruitful as a source of useful data than were the simpler response categories.

PROPOSED NEW METHOD FOR CONSTRUCTING
EMPIRICAL PERSONALITY SCALES WHICH
ARE WELL CONTROLLED FOR DIS-
SIMULATION TENDENCIES

Findings *d* and *e* above, however, did suggest a method for constructing empirically valid scoring keys for forced-choice personality inventories which would possess a rather high degree of control over the effects of attempted faking. What is more, this method should in general yield satisfactory keys *whether the stimuli in the forced-choice items are well matched on the faking stereotype or not*. In fact, we shall see that it is even a considerable advantage if there are sizable shifts in the response proportions under the two conditions—especially if the percentages of response under the straight-take condition are close to 50%!

Based on the data cited above, it was clear that if one could depend on the respondents to follow directions and give honest and accurate responses, Finding *d* implied that cross-validatable keys could be developed. Finding *e* implied that the proportion of per-

sons responding to each item in a given direction was extremely stable over different samples of persons under each of the two conditions taken separately. Thus, the *difference in popularity* of a response between the two conditions also would be a stable property of the items.

Now, to construct a scoring key for one of the criterion variables, suppose one were to initially select all the items on the test whose straight-take discrimination indices for the criterion factor were greater than some fixed amount, as he would do to build an ordinary empirical key. The actual number of items selected, however, should exceed the number of items wanted in the final key; the amount of the excess depending upon the number that will have to be discarded in the next step of the procedure. All items chosen, of course, should have statistically significant discrimination indices. The converse is probably not true, however, since, as will become apparent, one will want to maximize the number of items in the final key with straight-take popularity indices in the middle range where standard errors are larger.

Next, suppose one were to construct a distribution of these items on a continuum which runs from large negative differences in popularities through zero difference to large positive differences. The items which have a shift in popularity on the fake-take "away from" the keyed response (based on the straight-take data) will be plotted on one side of the zero point and those with shifts "toward" the keyed response on the other. Now if, as will generally be the case, the number of items on one side of the zero point is less than the number on the other, include all (or most) of the items from the less numerous set. Then begin to select additional items from the more numerous set, taking those with the largest discrimination indices first, *until the sum of the popularity-shift magnitudes on one side of the axis equals the sum on the other side*. All items with zero shifts might be added to the final key or not as desired. Scoring keys constructed in this manner will be hereafter referred to in this paper as Set III keys.

The first effect of this method, of course, is

to give a key on which the mean test score under the straight-take condition will equal that under the faking condition. This is so because the mean of the test score distribution (under each condition) is equal to the sum of the item popularities (under that condition), which have been equated by the procedures described in the last paragraph. Since these popularity indices under each of the two conditions (and consequently the differences between them to a slightly lesser degree) were found to be extremely stable over different samples of persons—Finding *e* above—very little disparity should exist between the means of the scores under the two conditions even for a cross-validation sample. As a result, the straight- and fake-take mean profiles across such a set of scales should be congruent and the problem encountered with the preliminary empirical scales for DAI, illustrated in Figure 1, accordingly should be solved. Since shifts in mean scores under different response conditions is likely to be a serious source of mistakes in selecting personnel, it is important to balance carefully the items whose popularity shifts are in the keyed direction against those for which popularity shifts are in the opposite direction—even though one thereby eliminates from the key some otherwise valid items.

A second desirable effect of this method will usually occur automatically (but could be maximized easily for any given set of data). Since a percentage difference discrimination index of item validity was proposed this is likely to result in a preponderance of items in the final key with straight-take popularity indices in the middle range.⁵ But if such

⁵ A percentage difference between median-split criterion highs and lows can attain its maximum value (100%) only when the overall popularity index is 50%—i.e., when all highs endorse the item and all lows do not. For popularities approaching either zero or 100%, the maximum percentage difference discrimination index attainable becomes progressively smaller. It is, of course, equal to zero at the popularity extremes whether the criterion groups are formed by a median split or by some other partition such as the approximate 30-40-30 percentages used here. A *t* ratio or probability associated with such a significance test statistic, if used as a validity index (as it has by some), would lead to the selection of more items with extreme popularities since the standard error of the per-

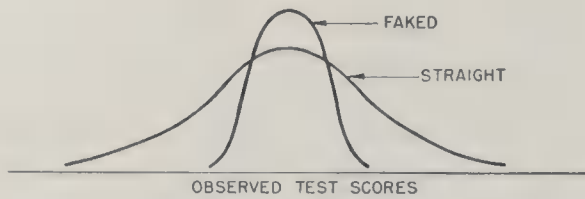


FIG. 2. Distributions of test scores for persons faking and those responding under straight self-report instructions for a hypothetical key of the sort described in the text.

items also have large fake-take shifts (in either direction), then *they will tend to have extreme fake-take popularities and correspondingly smaller item variances and inter-item covariances under the fake condition*. This means that the variance of the test scores on such a key under the fake-take condition will be smaller than that obtained under the straight-take condition. If in addition one elicits under the fake-take instructions inconsistent response sets across the items of the key, this should further lower the interitem covariances and further reduce the variance in test scores under this condition. The hypothesized effects of these item selection procedures are diagramed in Figure 2.

Thus, if one is primarily concerned with selecting persons who are extreme on the variable, the chances of his getting persons who have scored high or low on such a measure by attempting to fake is relatively small. The *degree* to which faking is controlled in this way, of course, will be a function of the difference in variances and the size of the selection ratio.

The second effect of the proposed item selection procedure is, of course, the reason behind the seemingly paradoxical statement made earlier; namely, that in order to get both good control on the contaminant and valid measures of the criteria, one *wants* items in the pool for which large shifts in response frequencies occur under attempted faking. This statement now should be qualified, however, to pertain to those items which have initial straight-take popularities in the middle range and have criterion validities of sufficient magnitude to be useful for inclusion in a scoring key.

centage difference becomes smaller as item popularities become more extreme.

These two qualifications suggest still one additional procedure that can be employed to achieve further control over the problem of faking. But here we want exactly the opposite pair of item characteristics to the two just described. If one can find an appreciable number of items which display large shifts in popularity under the two conditions but which have *extreme popularities* under self-report conditions and *near-zero validities* for all criterion variables, then the item is tailor-made for inclusion in a "detection" scale for identifying dissimulated test protocols. To build such a scale one simply keys for such items the responses which are infrequent under the straight-take condition. Scores on such a scale then indicate a performance on the test more or less similar to the modal faked performance, and if sufficiently high may constitute a basis for rejecting out of hand the test performance (and perhaps the respondent as well).

If the separation of straight and faked performances on such a detection scale is large and/or the selection ratio is high (hopefully, both), and if this scale is uncorrelated with the criteria and the measures of them, then its use can lower considerably the probability of selecting a "faker" without raising the probability of discarding valid highs on the content measures. But if one has to use all applicants, then elevated detection scores are troublesome since one then has *prima facie* evidence that the content scales he wishes to use are in all probability invalid for such a subject.

The arguments and evidence for constructing and using suppressor or correction scales to "adjust" content measures for the influence of response contaminants—usually test taking defensiveness or something similar—will not be discussed here. Suffice it to say that as long as the primary predictor is accounting for less than half the criterion variance—a situation universally true in personality assessment today and likely to persist for some time—one is well advised to attack the unexplained criterion variance rather than attempt to suppress invalid test variance. Statistically, a suppressor scale has to be a *very* effective suppressor—better than any

TABLE 4

PSYCHOMETRIC CHARACTERISTICS OF THE SET III KEYS FOR THE DAI, FCSRI-A, AND FCSRI-B AND THE SET III SUMMATION KEYS DEVELOPED ON SAMPLES A AND B AS ESTIMATED FROM THE RESPECTIVE CROSS-VALIDATION DATA

Sample A keys—Sample B data										Sample B keys—Sample A data										
Test	Key	Num- ber of items	Means		Standard deviations		Relia- bility		Straight			Num- ber of items	Means		Standard deviations		Relia- bility		Straight	
			S	F	S	F	K-R	20	C ₁	C ₂			S	F	S	F	K-R	20	C ₁	C ₂
DAI	I	56	26.5	27.4	9.56	7.00	.87	.87	.38	.46		52	25.2	24.2	7.59	4.98	.81	.81	.45	.51
	II	47	25.4	24.7	6.53	5.41	.77	.77	.36	.39		56	29.4	29.7	9.57	6.96	.88	.88	.26	.29
	III	54	29.2	28.2	9.02	6.72	.86	.86	.41	.41		53	27.3	28.7	8.26	5.54	.83	.83	.42	.49
	IV	51	24.7	24.3	5.17	3.79	.58	.58	.32	.40		54	27.3	27.3	6.00	4.52	.68	.68	.25	.33
	V	40	19.2	18.8	4.94	4.16	.64	.64	.33	.39		52	26.4	26.5	5.70	4.97	.65	.65	.30	.31
FCSRI-A	I	55	28.7	29.4	8.20	6.25	.83	.83	.41	.48		55	28.4	28.1	7.22	5.29	.77	.77	.51	.52
	II	56	26.4	26.2	5.58	5.21	.60	.60	.41	.39		53	25.0	25.5	7.32	5.06	.78	.78	.28	.36
	III	54	26.4	25.5	6.66	5.57	.73	.73	.47	.48		57	27.6	28.0	7.94	5.16	.81	.81	.39	.48
	IV	49	25.3	26.2	5.16	3.91	.59	.59	.28	.39		55	28.0	27.7	5.59	4.41	.60	.60	.21	.26
	V	54	25.1	25.6	5.01	4.16	.50	.50	.31	.34		49	23.7	23.9	5.99	4.64	.70	.70	.30	.36
FCSRI-B	I	54	28.4	29.3	7.42	5.28	.79	.79	.38	.44		51	26.7	26.2	6.47	4.72	.73	.73	.45	.44
	II	54	25.4	25.1	5.46	4.80	.59	.59	.22	.25		51	25.4	26.3	6.20	4.74	.70	.70	.14	.22
	III	50	25.4	25.1	6.88	5.18	.77	.77	.45	.45		53	26.6	27.3	6.91	4.95	.76	.76	.39	.49
	IV	55	26.8	27.4	4.91	4.18	.49	.49	.20	.34		50	25.8	25.4	5.35	4.71	.60	.60	.14	.25
	V	52	25.4	26.3	4.55	4.21	.41	.41	.26	.32		50	27.4	27.4	5.05	4.44	.56	.56	.24	.27
Summation	I	165	83.6	86.1	22.99	15.57	.93	.93	.43	.51		158	80.4	78.6	19.08	11.77	.91	.91	.53	.55
	II	157	77.3	75.9	14.78	12.45	.84	.84	.40	.42		160	79.9	81.6	21.01	13.93	.92	.92	.26	.32
	III	158	81.0	78.5	20.45	14.28	.92	.92	.49	.49		163	81.6	84.1	21.01	11.89	.92	.92	.44	.54
	IV	155	76.9	77.8	12.45	8.73	.78	.78	.33	.46		159	81.2	80.4	14.59	9.98	.83	.83	.23	.32
	V	146	69.9	70.6	11.84	9.41	.76	.76	.37	.43		151	77.3	78.0	14.16	10.58	.83	.83	.33	.37

Note.— $N = 228$ in each sample.

* Criterion C_1 = factor score from the peer nomination ratings obtained by simple summation of the scores obtained on the four salient scales for each factor. Criterion C_2 = C_1 minus the elevation component of the individual's factor score profile.

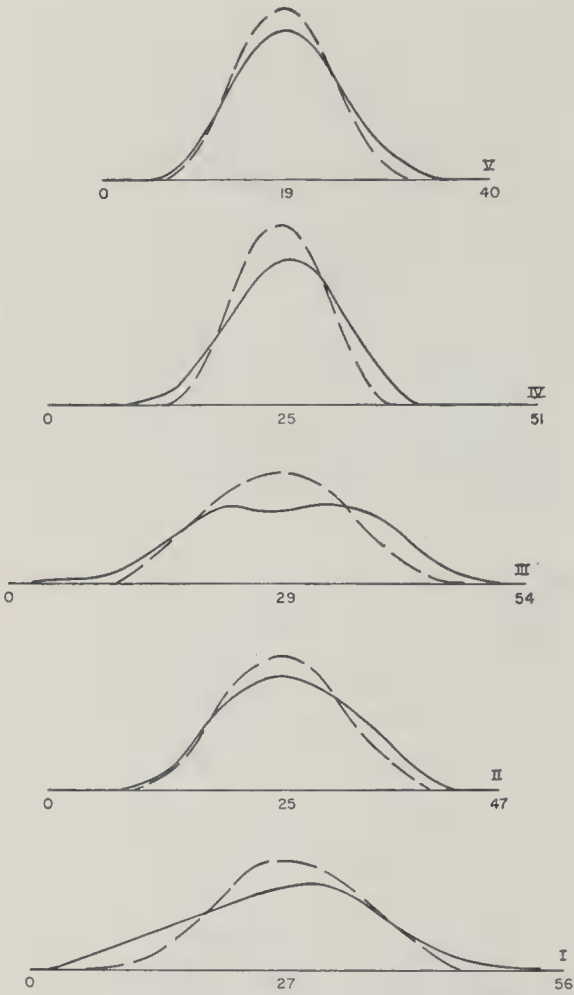


FIG. 3. Distributions of persons in Sample B on the Set III Sample A keys for the DAI under straight self-report (solid line) and faking (dashed line) conditions.

currently available or likely to be constructed soon—to effect any appreciable improvement over an only moderately valid primary predictor.

It is interesting to note again that in the construction and use of detection scales, just as in the development of the Set III type criterion predictor measures proposed above, the success of the attempt depends, at least in part, upon having items in the inventory which *do* show marked shifts in frequency of endorsement under the two administrative conditions. In the next section evidence will be presented that indicates that these methods of scale construction can be used effectively to construct valid measures of personality characteristics while simultaneously controlling for the influence of faking tendencies.

RESULTS OF THE DOUBLE CROSS-VALIDATION ANALYSES OF THE SET III KEYS AND DETECTION SCALES FOR DAI, FCSRI-A, AND FCSRI-B

Set III scoring keys were developed according to the procedure described above for each of the five criterion variables on each of the three tests. This was done separately on each of the double cross-validation Samples A and B, thus yielding two keys for each criterion factor on each test. The 15 keys based on each sample were then used to score both the straight and fake-take test performances by persons in the other (cross-validation) sample.

The extent to which the two desired faking-control properties of these keys were met is indicated in Figure 3 and in Table 4. In Figure 3 the distributions of Sample B scores on Sample A Set III keys for DAI are plotted. The medians of the two distributions for each scale are never more than two raw score points apart and in each case the range of the straight-take scores exceeds that for the fake-takes on both ends of the continuum. From Table 4 it can be seen that the means on all 15 Sample A keys for these subjects under the two conditions in no instance differ by more than one raw score point and in every case the standard deviations are greater under the straight-take condition. In the right half of Table 4 in the corresponding columns of means and standard deviations for the Sample B Set III keys estimated from Sample A data, essentially similar results may be found. In short, the method proposed does yield scoring keys for which mean profiles across the scales are, for all practical purposes, congruent and for which the straight self-report performances are more

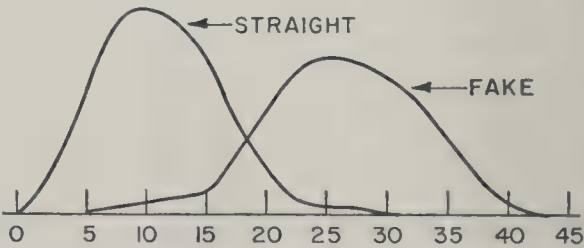


FIG. 4. Distributions of persons in Sample B on the Sample A Detection Key for the DAI under straight self-report and faking conditions.

TABLE 5

PSYCHOMETRIC CHARACTERISTICS OF THE DETECTION KEYS FOR THE DAI, FCSRI-A, AND FCSRI-B AND THE SUMMATION DETECTION KEYS DEVELOPED ON SAMPLES A AND B AS ESTIMATED FROM THE RESPECTIVE CROSS-VALIDATION DATA

Detection key for test	Sample A keys—Sample B data							Sample B keys—Sample A data						
	Num-ber of items	Means		Standard deviations		Reliabil-ities K-R 20		Num-ber of items	Means		Standard deviations		Reliabil-ities K-R 20	
		S ^a	F ^b	S	F	S	F		S	F	S	F	S	F
DAI	45	11.7	27.0	5.16	6.67	.70	.79	47	12.9	28.2	5.75	6.87	.74	.79
FCSRI-A	36	11.0	24.0	4.28	4.08	.61	.74	39	12.0	24.7	4.54	5.34	.62	.73
ACSRI-B	40	14.1	26.3	4.23	5.41	.53	.73	41	14.5	27.2	4.11	5.30	.48	.71
Summation	121	36.9	77.2	10.71	14.71	.79	.89	127	39.3	80.2	11.86	14.84	.82	.88

Note.—*N* = 228 in each sample.

^a S = straight, self-report administration.

^b F = attempted-fake administration.

variable than those obtained under attempts to fake a desirable test protocol.

In this latter regard, the differences in variability are not so large as one might wish. It is clear from the distributions presented in Figure 3 that for any but the most extreme cut scores, sizable proportions of fakers would be obtained. There are, however, several reasons why this does not constitute a serious deterrent to the use of these scales for personnel selection purposes.

First, and perhaps of only limited relevance, is the fact that separation between the scores under the two conditions is better at the low end of Scale I and at the high end of Scales II through V. This relative skewing of the straight-take scores, apparent in the results for DAI plotted in Figure 3, is exactly the pattern desired in our present situation where persons to be selected should possess high scores on Variables II-V and low scores for Factor I (Tupes, 1957). Why the distributions display this pattern of relative skewing is not entirely clear at the present time, but presumably it is a function of the distributions of discrimination indices and/or popularities for items chosen from the two sides of the popularity shift continua.

A second, and more general reason for satisfaction with the relative magnitude of the variances is that the effect is apparently cumulative across additional measures, at least for some of the variables. That is, if

one simply adds scores on Factor I scales for the three tests under each condition and plots these summation scores, the intersection of the two curves is *relatively* closer to the mean. For example, for Factor I scores on DAI, the two curves cross at a score of about 18. The proportion of straight-take scores less than 18 is .19, and the proportion of fake-takes in this same region is .08. For the summation score distributions on this variable the curves cross at a raw score of 68 and the corresponding proportions which scores less than this value are .28 and .09, respectively. The differences in means and standard deviations for the summation score distributions are given at the bottom of Table 4.

The final, and most poignant reason for unconcern about unwittingly getting fakers in the selection subsample is the effectiveness with which a faked answer sheet can be detected. In Figure 4, the Detection scale for DAI based on Sample A item analysis data has been applied to the two kinds of answer sheets for persons in Sample B. The separation is remarkably good. The false positive and false negative rates are each about 9% using the intersection point based on the original validation sample as a cut score.

Happily, one can even improve on this by using the summation of the detection scales across the three tests. In Table 5, the means, standard deviations, and K-R 20 reliabilities (all cross-validation sample estimates) for

the detection scales on each instrument and for the Summation Detection Scales are presented. The means for straight and faked scores on each test run about 2.5 to 3 *raw score* sigma units apart and for the Summation scales the separation is almost four times the standard deviation of the straight-take summation scores. The false positive and false negative rates for these Summation Keys are about 6%. Clearly if one can identify so large a proportion of faked answer sheets, he need not worry very much about getting such persons in his selection sample, even if such a person should happen to have a score on the content variable in the acceptable region.

Two additional features of the data in Table 5 are of interest before we return to the Set III content scale results. First, the standard deviations under straight-take are less than under the faking condition and secondly, the reliabilities under the faking condition are uniformly higher than under the self-report condition. Both of these effects are desirable. The first means that straight-take scores are not only lower but more tightly clustered about their mean value than are the faked scores. The second implies that for precisely those scores one wants to "interpret" (i.e., the faked ones) the reliabilities or interpretabilities (cf. Cronbach, 1951) are higher—in the case of the Summation Detection Scales, quite high.

It is of some importance to note that correlations between these detection scales with both the criterion factor scores and the Set III content scales are low. For the Sample A detection keys the highest value with any of the criterion ratings is a .16 between the Factor III and the DAI Detection scale. For the detection scales on the separate tests versus their respective Set III keys the highest values are $-.17$ (DAI, V), $-.24$ (FCSRI-A, II), and $-.34$ (FCSRI-B, II). For the Summation Detection Scale against the summation content scales the largest is a $-.34$ against Summation Scale II. It is also interesting to observe that the detection scales are rather uniformly negatively correlated with straight-take test scores on Variables II and III whereas they are correlated essentially zero with the criterion ratings on these factors. Thus, positive

weighting of the detection scale (reverse suppressor situation) could conceivably improve prediction of the criterion ratings. But, for reasons briefly noted earlier, the improvement would likely not be very large.

And finally, the median correlation among the pairs of separate Sample A Detection Scales under the straight-take condition is .41 whereas under the faking condition the median of the three correlations is .61. This supports the data in Table 5 where we noted that the internal consistency reliabilities (K-R 20s) were higher for the faked data.

Reference to the last three columns in each half of Table 4 provides an indication of the reliabilities and validities for the Set III keys. It can be seen that the K-R 20s for the separate scales are reasonably high for Scales I, II, and III for each inventory, generally lower for Scales IV and V. For the Summation scales, the reliabilities are really quite respectable, and for Variables I and III (and perhaps II, also), compare favorably with the best of available ability measures.

The validities in the first of the two columns (Criterion C_1) are correlations against the peer nomination rating factor scores; those in the other column (Criterion C_2) are the relationships against the rating factors with the elevation component of each individual's criterion score profile removed. That is, in this latter case, the validities are against criteria for which each subject has the same average profile across the criteria. Since, in general, elevated scores are "good" on these factors in terms of a general social desirability stereotype, the elevation component in the criterion ratings for a subject is a rough measure of his "esteem" or "eminence" or "likability" or some such general evaluative attribute as perceived by his peers. If one is not so much concerned about this kind of general characteristic as about the "pattern" of attributes on the set of factors and the ability of one's tests to predict such differential traits, then the values in the C_2 column are more pertinent. They are also almost uniformly higher.

Unfortunately, it is the C_1 criteria we set out to measure since it was for these rating variables that Tupes (1957) demonstrated predictive validity against his overall per-

formance criterion. However, even the C_1 validities for the scales on the separate inventories are all positive, all statistically significant, and only 6 of the 30 coefficients are less than .25. While these magnitudes are not high by absolute standards, they compare favorably with those for other personality inventories when independent, non-self-report criteria are used. In addition, since the selection program for which these measures are intended typically selects only about 10% of the applicants, these scales can probably be used with some guarded confidence. The validities in the lower part of Table 4 for the longer summation scales are somewhat better against both criteria as one might expect.

In constructing the Set III keys, we were more concerned with maximizing validities of these scales against the criterion ratings and in controlling against susceptibility to faking than we were in the relationships among scales. Consequently, we permitted the keying of a given item on more than one scale if the discrimination indices warranted it. The effect of this, of course, is to give us spurious interscale correlations due to correlated item error components. As a consequence, it will be well, for the present at least, to avoid trying to interpret relationships among our inventory scales in substantive psychological terms. Nor should they, in their present form, be used in factor analyses or other theoretically oriented multivariate studies.

Finally, it should be pointed out that the scores on the Set III keys under the fake-take condition correlate appreciably with *nothing* (except, possibly, other fake-take scores). The highest correlation of one of these scales under this condition with any criterion variable in either sample (a total of 300 coefficients) is a .20 (between C_2 Factor III and the FCSRI-A Sample A key for Variable IV, sic!). The correlations between the same keys applied under the two conditions are all less than .27. The corresponding-variable Set III keys across the three tests do intercorrelate to a moderate degree under the fake-take condition. This latter is another reflection of a certain stability or consistency of the faking stereotypes of the persons in our

samples previously noted in the intertest detection scale correlations.

Thus, it should be clear that if a person assumes a faking attitude comparable to the modal stereotype displayed by our subjects, psychological interpretations of his personality attributes on the basis of his Set III key scores on these tests is a fool's enterprise. Fakers of this sort, however, should be very easy to identify from their detection scale scores.

SUMMARY, DISCUSSION, AND SOME IMPLICATIONS OF THESE FINDINGS

The results presented in the last section indicate that if one constructs self-report inventory measures of personality characteristics by the method proposed in this paper using content-relevant stimuli and binary forced-choice item formats, then the following results can be achieved on cross-validation data:

1. Validities against peer nomination, rating criteria which are in the moderate range and which are slightly, but uniformly higher for criteria from which profile elevations have been removed.

2. Internal consistency estimates of reliability ($K-R$ 20) which are moderate to high for the individual scales. For the longer summation scales these are uniformly over .75 and in the case of two (or possibly three), of the five variables, greater than .90.

3. Very tight control over the problem of faking as reflected in: (a) mean profiles across the scales which are congruent under straight, self-report conditions and under instructions to attempt to fake; (b) variances for the faked answer sheets which are smaller than for the straight-take performances; and (c) detection scales which, for a single inventory key, identify about 91% of the fakers at a cost of misclassifying only about 9% of the nonfakers. With the longer Summation Detection Scales the false positive and false negative rates drop to about 6%.

With regard to this latter point, it is conceivable that little, if any, further improvement can be made in the detection of faked answer sheets. In the case of the false positives (i.e., those who get high detection scale

scores under the straight-take conditions), it is just possible that they actually are like the modal faking stereotype in their personality make-up. It is also possible, of course, that for whatever reasons, these persons assumed a faking attitude under the initial straight self-report instructions. In the case of the false negatives (i.e., those with low detection scale scores under the faking condition), it could be either that they have markedly disparate notions of what constitutes a desirable set of test responses, that they misunderstood the instructions, or that they assumed that the best way to "fake one's way into OCS" was to be as "honest and accurate as possible." Whatever the case, a cursory inspection of the detection scale scores for these two kinds of cases reveals that, in most instances, the scores under the two conditions for such persons are highly similar. No follow up of these cases by interview or other means was attempted, but this might have proved to be most interesting and quite feasible had we only anticipated how few cases we were going to miss with our detection scales.

By way of general evaluation of these findings it is clear that inventory measures constructed in the manner here proposed can be used with considerable confidence for personnel selection where one is interested in assessments of status on these criterion dimensions—at least in situations where the selection ratios are fairly small. In any event, the probabilities of mistaking intentional fakers as "honest and accurate" self-reporters are very low and, if all three inventories are used, all but negligible.

A word of caution is in order, however. A preliminary study based on a small sample of Peace Corps trainees who were asked to fake a different desirability stereotype (admission to Peace Corps volunteer status) after first taking DAI and FCSRI-A under standard straight take instructions, yielded mean profiles on the original (Sample A) Set III keys which under the two conditions, were markedly noncongruent. Nor did the original detection scales described above separate very adequately the straight-take from the faked performances by these subjects. Inspection of these data revealed that the straight-take mean

profile on the Set III keys was only slightly different than that obtained from the subjects in the original cross-validation sample. But there were marked disparities between the profiles from these two groups based on their respective faked performances. Clearly the modal stereotype of what was seen as desirable by college males in order to gain admission to a military officer training school was quite different (and in ways that are not very surprising) from what was viewed as desirable by Peace Corps trainees in order to be accepted as a volunteer in that organization.

One implication of these findings is (unfortunately) quite clear. Control over faking of the sort achieved by Set III keys and detection scales, constructed for use in one setting with one class of respondents, may not generalize very widely. Just what the limits of generalizability attainable may be is as yet not fully known. But it does seem to be highly improbable that it will be possible to construct and standardize a single set of general purpose keys which can be used effectively in all settings for all classes of respondents. It may even prove to be necessary to build special purpose keys for each new context where they are desired. This is not a very happy prospect, but it is one that we may have to learn to live with all the same.

REFERENCES

- BERKSHIRE, J. R. Comparisons of five forced-choice indices. *Educ. psychol. Measmt.*, 1958, **18**, 553-561.
- BORISLOW, B. The Edwards Personal Preference Schedule (EPPS) and fakability. *J. appl. Psychol.*, 1958, **42**, 22-27.
- CATTELL, R. B., Confirmation and clarification of primary personality factors. *Psychometrika*, 1947, **12**, 197-220.
- CORAH, N. L., FELDMAN, M. J., COHEN, I. S., GRUEN, W., MEADOW, A., & RINGWALL, E. A. Social desirability as a variable in the Edwards Personal Preference Schedule. *J. consult. Psychol.*, 1958, **22**, 70-72.
- CRONBACH, L. J. Response sets and test validity. *Educ. psychol. Measmt.*, 1946, **6**, 475-494.
- CRONBACH, L. J. Further evidence on response sets and test design. *Educ. psychol. Measmt.*, 1950, **10**, 3-31.
- CRONBACH, L. J. Coefficient alpha and the internal structure of tests. *Psychometrika*, 1951, **16**, 297-334.

- CRONBACH, L. J. Assessment of individual differences. In P. R. Farnsworth (Ed.), *Annu. Rev. Psychol.*, 1956, 7, 173-196.
- DUNNETTE, M. D., MCCARTNEY, J., CARLSON, H. C., & KIRCHNER, W. K. A study of faking behavior on a forced-choice self-description checklist. *Personnel Psychol.*, 1962, 15, 13-24.
- EDWARDS, A. L. *Manual for the Edwards Personal Preference Schedule*. New York: Psychological Corporation, 1954.
- EDWARDS, A. L. *The social desirability variable in personality assessment and research*. New York: Dryden Press, 1957.
- EDWARDS, A. L., WRIGHT, C. E., & LUNNEBORG, C. E. A note on "Social desirability as a variable in the Edwards Personal Preference Schedule." *J. consult. Psychol.*, 1959, 23, 558.
- GUILFORD, J. P. *Psychometric methods*. (2nd ed.), New York: McGraw-Hill, 1954.
- JURGENSEN, C. E. Report on the "Classification Inventory," a personality test for industrial use. *J. appl. Psychol.*, 1944, 28, 445-460.
- KELLY, E. L., & FISKE, D. W. *The prediction of performance in clinical psychology*. Ann Arbor: Univer. Michigan Press, 1951.
- LOEVINGER, J. Theory and techniques of assessment. In P. R. Farnsworth (Ed.), *Annu. Rev. Psychol.*, 1959, 10, 287-316.
- LONGSTAFF, H. P., & JURGENSEN, C. E. Fakability of the Jurgensen Classification Inventory. *J. appl. Psychol.*, 1953, 37, 86-89.
- MAHER, H. Studies of transparency in forced-choice scales: I. Evidence of transparency. *J. appl. Psychol.*, 1959, 43, 275-278.
- MESSICK, S. Dimension of social desirability. *J. consult. Psychol.*, 1960, 24, 279-287.
- NORMAN, W. T. Development of self-report tests to measure personality factors identified from peer nominations. *USAF ASD tech. Note*, 1961, No. 61-44. (a)
- NORMAN, W. T. Problems of response contamination in personality assessment. *USAF ASD tech. Note*, 1961, No. 61-43. (b)
- NORMAN, W. T. Toward an adequate taxonomy of personality attributes: Replicated factor structure in peer nomination personality ratings. *J. abnorm. soc. Psychol.*, 1963, 66, 574-583.
- ROSEN, E. Self-appraisal, personal desirability, and perceived social desirability of personality traits. *J. abnorm. soc. Psychol.*, 1956, 52, 151-158.
- TUPES, E. C. Relationships between behavior trait ratings by peers and later officer performance of USAF Officer Candidate School graduates. *USAF PTRC tech. Note*, 1957, No. 57-125.
- TUPES, E. C., & CHRISTAL, R. E. Stability of personality trait rating factors obtained under diverse conditions. *USAF WADC tech. Note*, 1958, No. 58-61.
- TUPES, E. C., & CHRISTAL, R. E. Recurrent personality factors based on trait ratings. *USAF ASD tech. Rep.*, 1961, No. 61-97.
- WHERRY, R. J., STANDER, N. E., & HOPKINS, J. J. Behavior trait ratings by peers and references. *USAF WADC tech. Rep.* 1959, No. 59-360.

(Received September 10, 1962)

SELF-PERCEPT AND CONSUMER ATTITUDES TOWARD SMALL CARS

EUGENE JACOBSON AND JEROME KOSSOFF¹

Michigan State University

A random household sample of 250 adults in a metropolitan area was interviewed to determine the relationship between respondents' attitudes toward the purchase of small cars and readiness to accept challenge and innovation. It was expected that persons who saw themselves as Confident Explorers would show greater acceptance of small cars. The data did not support this relationship. On the contrary, persons who saw themselves as Cautious Conservatives were more likely to express a favorable opinion of small cars. Conservatives did not consider the small car an adventure, but rather a practical, economical convenience. Confident Explorers saw a large car as a means of expressing their ability to control the environment.

One of the important new objects of consumer behavior that has appeared in the past few years is the smaller automobile. It has emerged in the United States with relative suddenness and is the occasion for discussion, expressed preferences, and crystallization of attitudes.

As an innovation, it is of interest to the student of consumer behavior from a number of points of view. Among these is the possibility of developing measures for relating a readiness to accept small cars with a readiness to accept other kinds of innovation. Perhaps some people are more likely to express favorable attitudes toward small cars simply because, in general, they consider themselves to be the kind of people who are ready to try something new and different.

In the study reported in the following paragraphs, an attitude questionnaire was administered to a random sample of adults in a metropolitan area to discover whether persons who consider themselves ready for challenge and innovation are also persons who express more positive attitudes toward the purchase of small cars. The findings were that this was not the case but rather that those persons who saw themselves as being cautious and conservative were more likely to express a positive attitude toward the purchase of small cars. The data provides some under-

standing of why this alternative relationship was found.

METHOD AND SAMPLE

The population sample consisted of 250 adults, all of the persons 21 years old or older in 116 households chosen at random from a predesignated area in Woodside, Long Island, New York. Interviewing took place from June through September 1960.

As a measure of attitudes toward the self, respondents were asked to complete a 16-item Likert-type scale. The two items reproduced here are representative of the scale:

"Unless there is a good reason for changing, I think we should continue to do things the way they are being done now."

1. Agree Very Much
2. Agree Somewhat
3. Neutral
4. Disagree Somewhat
5. Disagree Very Much

"When new ideas are going around, I am usually among the first to accept them."

1. Agree Very Much
2. Agree Somewhat
3. Neutral
4. Disagree Somewhat
5. Disagree Very Much

On the basis of total scores, cutting points were established and three groups were identified and labeled as Cautious Conservatives, Middle-of-the-Roaders, and Confident Explorers. The total scale had a Spearman-Brown corrected estimated internal consistency of .84.

In comparing the three groups, the Cautious Conservatives were more likely than the others to give responses indicating that they saw themselves as unwilling to take chances, wanting to have proved

¹This report is based on an MA thesis submitted to the Department of Psychology, Michigan State University, by Kossoff, in 1961.

methods, and preferring safety to adventure. Confident Explorers were more likely to see themselves as ready to challenge the unknown, to trust their ability to handle situations, and to be free in the use of their resources.

RESULTS

All three groups expressed more favorable than unfavorable attitudes toward small cars. A significantly larger percentage of the Confident Explorers expressed unfavorable attitudes toward small cars, and correspondingly, Cautious Conservatives were more likely to be favorable (see Table 1).

The same relationship was found in preferences for a second car. Confident Explorers were more likely to want a big car, Cautious Conservatives to prefer a small car.

TABLE 1
SELF-PERCEPT AND ATTITUDE
TOWARD SMALL CARS

Attitude	Cautious conserv- atives	Middle- of-the- roaders	Confident explorers
Favorable	79	58	49
Neutral	10	8	9
Unfavorable	10	34	42
Total %	99	100	100
<i>N</i>	86	71	93
$\chi^2 = 22.51^*$			

Note.—Total *N* = 250.
* $p < .01$.

When we examine correlates of these attitudes, several factors having a bearing on car preference emerge. One important relationship is found in the difference in attitudes between men and women. Women are much more likely to have a favorable attitude toward small cars than men (see Table 2).

This led us to re-examine the data on attitude toward the self where the data showed that women were more likely to regard themselves as cautious and conservative than men (see Table 3).

However, when we examined attitudes toward small cars, self-percept, and sex simultaneously, the importance of self-percept was confirmed. As would be expected, most of the women who saw themselves as Cautious

TABLE 2
SEX AND ATTITUDE TOWARD SMALL CARS

Attitude	Men	Women
Favorable	48	77
Neutral	8	10
Unfavorable	43	14
Total (%)	99	101
<i>N</i>	130	120
$\chi^2 = 27.08^*$		

Note.—Total *N* = 250.
* $p < .01$.

Conservatives were favorable toward small cars. But the Cautious Conservative men had the same pattern of response (see Table 4).

Although there were only a handful of women who had unfavorable attitudes toward small cars, those who did were Confident Explorers or Middle-of-the-Roaders. Confident Explorer men were more likely than Cautious Conservative men to have an unfavorable attitude.

This general pattern of relationships is illustrated in a second analysis, that of attitudes toward American-made small cars as opposed to foreign-made small cars. Cautious Conservatives are more likely to prefer American small cars, and Confident Explorers tend to prefer foreign small cars. Women prefer American small cars.

Women's attitudes in this area do not differ significantly from age group to age group. However, among men there are some interest-

TABLE 3
SELF-PERCEPT AND SEX

Sex	Cautious conserv- atives	Middle- of-the- roaders	Confident explorers
Men	40	48	67
Women	60	52	33
Total (%)	100	100	100
<i>N</i>	86	71	93
$\chi^2 = 15.1^*$			

Note.—Total *N* = 250.
* $p < .01$.

TABLE 4
SELF-PERCEPT, ATTITUDE TOWARD SMALL CARS AND SEX

Attitude	Cautious conservatives		Middle-of-the-roaders		Confident explorers	
	Men	Women	Men	Women	Men	Women
Favorable	71	85	44	70	39	71
Neutral	6	13	9	8	10	6
Unfavorable	23	2	47	22	51	23
Total (%)	100	100	100	100	100	100
N	34	52	34	37	62	31
	Men: $\chi^2 = 8.7^*$		Women: $\chi^2 = 10.20^*$			

Note.—Total N = 250.
* $p < .01$.

ing variations. Men in the age group 20–29 show the most marked interest in small foreign cars. The strongest preferences for big cars among men is found in the 40–49 age group.

Among women, differences in education have relatively little bearing on attitudes toward themselves or small cars. Among men, those who have more education are more likely to prefer a foreign small car.

Income is a factor in car preference among men. Those reporting less than \$5,000 a year income are more likely to prefer small American cars. Those in the higher income brackets are more likely to prefer big cars or foreign-made cars if a small type is considered.

Of the 130 men interviewed, 112 owned cars. Ninety-four owned big cars; 14, small cars; and 4, sports cars. All of the small car owners had a favorable attitude toward small cars and a preference for American makes. Among the large car owners, most preferred large cars and foreign small cars. Among the 120 women interviewed only 36 owned cars, 29 large and 7 small. Those who owned big cars were much more likely to be favorable toward the small car than their big-car-owning male counterparts.

In our sample, Cautious Conservatives were likely to be in their late forties or early fifties with a high school education and an income in the \$5,000 range.

Confident Explorers as a group are younger.

Sixty-six percent are in the 20–39 age group as compared to 8% of the Cautious Conservatives. But, within the Confident Explorer group, the men and women who prefer big cars are, on the average, 6 years older.

Confident Explorers, both men and women, are better educated than Cautious Conservatives and have a slightly higher income. They are likely to have a college degree or some college training. Highest income was reported by the male Confident Explorers who preferred large cars. This subgroup, upon further analysis of the components of the attitude scale, was found to have the most liberal attitudes toward the use of money.

DISCUSSION

A coherent pattern of attitudes toward small cars can be constructed from these findings.

First it is clear that there is a strong and meaningful relationship between expressed attitudes toward the self and expressed preferences among cars. The relationship, however, is not as simple as the one predicted. Apparently the Cautious Conservative prefers the small American car, accepting it as a recognizable, more convenient, scaled-down version of the vehicle with which she or he is already familiar rather than a challenging innovation.

There seem to be two major kinds of Confident Explorers. First there is the relatively

high income, not so young, fairly well-educated man who expresses a preference for the standard large American car. Although he can be challenged by opportunity for innovation and feels competent to handle new experiences, he does not react to the small American car as a challenging innovation. Rather he sees it as a low priced less comfortable vehicle and he feels that he can afford to own a large car with more conveniences. Part of his attitude of confidence is based on his access to money that he can use freely.

Then there is the relatively young, relatively well-educated man or woman who prefers the small car. These people, who see themselves as Confident Explorers too, are perhaps the model the authors had in mind

when they began the study. Although in relatively high income brackets, for most of them money is committed to basic family expenses. For them the small car is a creative alternative to more expensive standard size cars. But also in this group are the individuals with relatively large amounts of disposable income who prefer sports cars, foreign and domestic.

Predictions about the Cautious Conservatives, then, can be made with fair probability of accuracy. It is highly likely that they will choose small American cars. Predictions about Confident Explorers have to be conditioned by additional information about income, age, sex, family status, and education.

(Received April 19, 1962)

UNDERLYING SOURCES OF JOB SATISFACTION¹

FRANK FRIEDLANDER²

Psychological Research Services, Western Reserve University

Recent theories of job satisfaction generally assume 2 underlying types of job elements important to employee satisfaction: those in the work process which allow for self-actualization, and environmental elements in which the worker's rewards are physical and monetary. A parallel assumption pertains to the 2 types of employees for whom each of these is important. A validation of such constructs was attempted through factor analysis and indicated 3 underlying groups of job elements important to job satisfaction: social and technical environment, intrinsic work aspects, and recognition through advancement. The factor of greatest import to each employee was identified, and factored groups of employees were described in terms of their differing age, salary, and occupational patterns. No significant differences in overall job satisfaction among the 3 groups were found.

Growing numbers of studies are concerned with an examination of the underlying sources of job satisfaction that are available to the worker in his job environment. A parallel question concerns the establishment and identification of categories of workers to whom specific features of the job environment are of greatest import. If such categories do exist, what are some of the correlative differences in the type of work, salary, and age of these groups? Finally, do such groups differ in the satisfactions which they find in whatever they might seek? The purpose of this study is to investigate these questions by analyzing attitude scales designed to measure the sources of satisfaction for several kinds of employees.

The implications of such problems are relevant to findings of a number of other studies. While acknowledging that the worker has a sense of attachment to his work and workplace, Dubin (1956) found a lack of a corresponding sense of total commitment to it. Weiss and Kahn (1959) found that a vast majority of workers view work as imposed, not enjoyed, and negative rather than freely chosen, productive, and positive. Argyris (1957) and McGregor (1960) deplore the

lack of need-satisfying and self-actualizing job elements available to the worker in an industrial environment which management has saturated with physical and security offerings. By implication, these writers assume a two-factor job context.

Several studies have indicated differences in satisfaction derived from intrinsic as opposed to extrinsic job factors. These were generally not based upon a clear delineation between the two types of job factors, but rather upon apparent differences in overall job motivation between the two groups of employees. Thus differences in motivation and morale were observed, and efforts were made to identify the causal determinants of these in terms of the intrinsic-extrinsic dichotomy. Gurin, Veroff, and Feld (1960) found that a greater satisfaction was derived from "ego satisfying" work, and a more limited and less intensive satisfaction when gratification comes mainly from extrinsic aspects of the job. Pelz (1957) found intrinsic motivations more essential for high research performance. Hoffman and Mann (1956) found "increased job interest derives from actually attacking the job content rather than altering the more peripheral aspects of the job situation." Kornhauser (1962) found workers in occupations in which well-developed skills are applied show greater mental health than those limited to repetitive machine-paced operations. In one of the few studies aimed directly at the intrinsic-extrinsic dichotomy, Super (1962) found no

¹ This paper is based upon a dissertation submitted to Western Reserve University in partial fulfillment of the requirements of the PhD degree. The author gratefully acknowledges the help of Joel T. Campbell under whose direction the research was conducted.

² Now at United States Naval Ordnance Test Station, China Lake, California.

evidence for such dual constructs as they related to work values.

Although the intrinsic and extrinsic factors have not been empirically established as independent factor structures, either as elements within the job context or as two distinct types of motivation, several investigators have used this dichotomy for relating job elements with job attitudes. Herzberg, Mausner, and Snyderman (1959) found satisfaction resulted primarily from intrinsic job elements while dissatisfaction was derived from the extrinsic elements. Similar findings were obtained by Hahn (1959) and Schwarz (1959); this latter, however, found that incidents leading to negative job attitudes usually involved frustrating managers' attempts at self-actualization. Friedlander (unpublished) found that, for the most part, job items important to satisfaction were uncorrelated with items important to dissatisfaction.

The distinction between extrinsic and intrinsic factors as differing sources of job satisfaction has thus emerged. From the few studies designed to pinpoint the elements of this dichotomy, constructs are evolving which help delineate and define their meaning. The importance of such a dichotomy cannot be denied; the validity of such constructs remains to be measured.

The purpose of this study, then, was to examine a number of the problems raised in the previous discussion. Specifically, this study was concerned with: an analysis of the elements within a job context so as to obtain construct validation of the underlying sources of job satisfaction, identification and description of employees for whom each group of job factors is of greatest importance as a source of satisfaction, and an analysis of differences in overall satisfaction among the groups of employees.

PROCEDURE

This study employed a questionnaire which was administered in October 1960 by a large midwestern manufacturing company to its engineering, supervisory, and salaried employees. Of the approximately 10,000 employees, over 92% completed the questionnaire. Two hundred of each of the three position-occupation groups were selected in random fashion from the entire group. The age of the members of

this sample was normally distributed from under 25 to over 55, with a mean of 39. The monthly base salary was of approximate normal distribution, ranging from under \$500 to over \$950, with a mean of \$738.

Thirty-nine questions were used to form three separate measures or tests: 17 of these measured the importance of various items to the employee's source of satisfaction, an additional 17 measured the actual satisfaction of the employee with the same items, and 5 separate items measured the overall satisfaction of the employee. In the first of these, the respondent was asked to think of a time he felt exceptionally good about any job that he had held. He then indicated (by checking one of four blanks) the extent to which each of the 17 items had contributed to this good feeling in the particular satisfying experience which he had considered.

Pearson product-moment correlations were computed among the 17 source-of-satisfaction items. The resultant matrix was factor analyzed by the principal-factor method. Squared multiple-correlation coefficients were used as approximations to the communalities.

All employees in the sample of 600 were then assigned three separate factor scores. All items with loadings above .30 on a factor were considered as making up that factor. Factor scores were transformed into standard scores, and a mean standard score was then computed for each employee, as well as the amount that each of these three standard scores deviated from his mean standard score. These deviation scores were then utilized to identify those 100 employees who had the highest possible deviation from their mean standard score in each of the three factors. In effect, those who indicated a specific factor as a great source of satisfaction relative to other employees were designated as representing that factor.

Whereas the previously discussed set of questions tapped the importance of various items as a source of job satisfaction, a second set tapped the extent to which the respondent perceived the availability of the same items in his present job environment. For each member of each factor group, satisfaction with the factor which he seeks was computed from his responses to the items in the second set of questions. This satisfaction score was computed in the same manner as his source-of-satisfaction score; the respondent's responses within each factor were summed. These scores were then used as a screening device for selecting those of the 300 respondents used in the final analysis.

The design was one of testing the significance of the differences in overall satisfaction among those in each of the three groups who were most satisfied with that which they sought. Lack of availability of the satisfiers was thus minimized since only those members within each of the three groups who were most satisfied with that which they sought were selected for comparison in overall satisfaction. A set of five questions which tapped the respondent's overall general satisfaction with his job and his company was used in this final analysis.

RESULTS

The correlations among each of the 17 sources of satisfaction as well as the rotated factor loadings have been deposited with the American Documentation Institute (ADI).³ Three meaningful factors emerged which were identified as follows:

Factor I—Social and Technical Environment:

- 4. The working relationship I had with my supervisor was very good .66
- 6. I was working under a supervisor who really knew his job .66
- 15. I was working in a group that operated very smoothly and efficiently .52
- 16. Management policies that affected my work group took into consideration the personal feelings of the employees .46
- 9. I had exceptionally good working conditions and equipment .39
- 11. I felt secure in my job .38
- 5. The working relationship I had with my co-workers at my level was very good .33

This factor encompasses the social and technical aspects of supervision, of the work group, and of the working conditions as a source of satisfaction. With the possible exception of Items 9 and 11, all items within this factor seem to provide interpersonal and “other-directed” (Riesman, Glazer, & Denney, 1951) sources of job satisfaction and job orientation.

Factor II—Intrinsic Self-Actualizing Work Aspects:

- 17. The job required the use of my best abilities .59
- 8. I had a real feeling of achievement in the work I was doing .56
- 13. I liked the kind of work I was doing .50

³ The correlation matrix, the rotated factor loadings, and the three chi square tables referred to later have been deposited with the American Documentation Institute. Order Document No. 7502 from ADI Auxiliary Publications Project, Photoduplication Service, Library of Congress, Washington 25, D. C. Remit in advance \$1.25 for microfilm or \$1.25 for photocopies and make checks payable to: Chief, Photoduplication Service, Library of Congress.

- 2. I received a particularly challenging assignment .48
- 12. I was getting training and experience on the job that were helping my growth .44

The development and full use of one’s capacities and talents seem to be central to most items in this factor. More specifically, each item seems to be related to the intrinsic work process itself and to the relationship of this process to the development and growth of the individual. The inclusion of Item 8, feeling of achievement, in this factor rather than in Factor III implies that achievement as a source of job satisfaction is more related to the successful development and use of one’s abilities than to the successful attainment of the more recognizable criteria of job ascendancy.

Factor III—Recognition through Advancement:

- 1. I felt there was a good chance that I’d be promoted .60
- 7. I was expecting (or received) a merit increase .48
- 2. I received a particularly challenging assignment .40
- 10. I was given increased responsibility in my job .40
- 3. A job I did received recognition as being a particularly good piece of work .34

Most items in this factor are concerned with recognizable signs of achievement as a source of job satisfaction. This factor also encompasses the challenging assignments and increased responsibility that generally accompany tangible evidence of recognition, such as an increased salary and advancement. When compared to the interpersonal elements within Factor I and the intrapersonal elements in Factor II, the items in Factor III appear to be of a more impersonal nature directed primarily toward ascendant strivings within the organization.

So as to obtain a better description of the respondents within each of the three factor groups, their characteristics among various age, salary, and position-occupation categories were analyzed. For this purpose, the three sets of factor scores were viewed as

the results of three separate factored tests. Kuder-Richardson reliabilities of these were .82, .72, and .70, respectively, for Factor I, Factor II, and Factor III.

Chi square tests³ indicated significant relationship between the source-of-satisfaction factors and the age, salary, and position-occupation categories. Those who derive satisfaction from the social and technical environment (Factor I) may be described as older, but less well paid, and were more frequently found in the salaried and supervisory groups; there was a significantly low frequency of these employees among engineers. A further analysis indicated there was much less of a positive relationship between age and salary in Factor I than in the other factor groups.

It might be surmised that the membership of the Factor I group consists of individuals who reached a lower level of education and who have had relatively slow progress in the organization. Furthermore, these individuals possess a rather strong need for good supervision and group relations. Employees who show this need may desire to defer to the leadership of supervision, management, or the work group. They also would seem to place importance on being able to rely upon others and thereby gain a measure of security. They are relatively unconcerned with promotion, challenge, and the kind of work they are doing.

Those in the Factor II group, who place prime importance on the more intrinsic job aspects which afford an opportunity for self-actualization, were found more frequently in the younger age groups. This finding would tend to indicate that those in younger age groups are more concerned with meaningful work which utilizes the best of their abilities and in which they might have a feeling of achievement. Perhaps they do not yet have great financial responsibilities so as to be concerned with salary and security. A greater proportion of their vocational career lies in the future so as to afford them the opportunity of being concerned with the more self-actualizing job aspects of gaining training and experience so as to grow.

No specific characteristics which differed from expectations were found in the group which derived satisfaction from recognition

through advancing in the organization (Factor III).

Finally, measures of overall job satisfaction among the three-factor groups were compared. The reliability, as computed by K-R 20, of the overall satisfaction measure was .73. An *F* test among these three means indicated no significant differences. Thus, there is no indication that any one of the three-factor groups had greater overall job satisfaction.

DISCUSSION

The results of this study in part substantiate and in part contradict a number of studies mentioned previously which have concluded with either the implication or indication of a two-factor structure of job satisfaction. Specifically, whereas Super (1962) found low or negative correlations between intrinsic items, this study found all such correlations significant beyond the .01 level, with a mean correlation of .32. The differences might well be due to the two separate approaches to the problems, i.e., values versus sources of satisfaction.

The study by Herzberg et al. (1959) relates to the current study since the job items extracted from interview protocols were all represented in the questionnaire utilized in the current study. Since those questions which dealt with sources of job satisfaction in the study by Herzberg elicited responses concerned primarily with intrinsic and growth aspects, one might expect a general intrinsic work aspect factor to emerge as the dominant one from a factor analysis of questions dealing only with satisfactions. The results of the current study tend to indicate that the underlying structure of job satisfaction is somewhat more complex than this.

Both intrinsic and extrinsic job factors were found as sources of job satisfaction. A factor very similar to Herzberg's "hygiene factor" emerged from the *satisfaction* sphere. Such results, however, were perhaps fostered by the mere inclusion of hygiene items in the questionnaire.

Although the study by Herzberg does not claim to have established empirically two independent and unique factors, some evidence is offered that job elements occur

together more frequently. Such relationships are in general accord with the results of the current study. Factors I and II in the current study correspond, in part, with Herzberg's concept of hygiene and motivation, although Factor III seems to draw from both the hygiene factor (merit increases) and the motivation factor (promotion and recognition).

In a similar study by Schwarz (1959), responses to a satisfaction question fell into two principal components: one dealt with competence, i.e., success of personal action, approach, or solution. The reinforcing agent in this component was a tangible indication of productivity or success. This component is similar to Factor II in the current study. A second component of the satisfiers was recognition or commendation for personal or group effort, and, to a lesser extent, concern with merit increases in pay. This component would seem to coincide with Factor III, Recognition through Advancement, in the current study.

Perhaps the factorial results of this study most clearly coincide with those suggested by Ginzberg, Ginsburg, Axelrad, and Herma (1951) who found that the individual usually recognizes three distinct, though related, types of satisfactions to be derived from work: the return in the form of monetary rewards and prestige; intrinsic satisfactions or the pleasure in a specific activity and in the accomplishments of specific ends; and concomitant satisfactions, such as those derived from working in a particular physical environment or with a particular group.

This study has thus identified three underlying dimensions within the sphere of satisfactions, and then ascertained the point on each of these three dimensions at which each

employee lies. Those employees at the extremes of each dimension view contrasting job elements as differing in importance as a source of satisfaction. They have been given further description in terms of differing age, salary, and occupational patterns.

REFERENCES

- ARGYRIS, C. *Personality and the organization*. New York: Harper, 1957.
- DUBIN, R. Industrial workers' worlds. *Soc. Probl.*, 1956, 3, 131-142.
- GINZBERG, E., GINSBURG, S. W., AXELRAD, S., & HERMA, J. L. *Occupational choice*. New York: Columbia Univer. Press, 1951.
- GURIN, G., VEROFF, J., & FELD, SHEILA. *Americans view their mental health*. New York: Basic Books, 1960.
- HAHN, C. P. *Dimensions of job satisfaction and career motivation*. Pittsburgh, Pa.: American Institute for Research, 1959.
- HERZBERG, F., MAUSNER, B., & SNYDERMAN, BARBARA B. *The motivation to work*. New York: Wiley, 1959.
- HOFFMAN, L. R., & MANN, F. C. Individual and organizational correlates of automation. *J. soc. Issues*, 1956, 9, 3-26.
- KORNHAUSER, A. Mental health of factory workers. *Hum. Organiz.*, 1962, 21, 43-47.
- MCGREGOR, D. M. *The human side of enterprise*. New York: McGraw-Hill, 1960.
- PELZ, D. C. *Motivation of the engineering and research specialist*. (General Management Series No. 186) New York: American Management Association, 1957.
- RIESMAN, D., GLAZER, N., & DENNEY, R. *The lonely crowd*. New Haven: Yale Univer. Press, 1951.
- SCHWARZ, P. A. *Attitudes of middle-management personnel*. Pittsburgh, Pa.: American Institute for Research, 1959.
- SUPER, D. E. The structure of work values in relation to status, achievement, interests, and adjustment. *J. appl. Psychol.*, 1962, 46, 231-239.
- WEISS, R. S., & KAHN, R. L. *On the definition of work among American men*. Ann Arbor: University of Michigan, Survey Research Center, 1959.

(Received July 26, 1962)

A NOTE ON *THE* CRITERION

MARVIN D. DUNNETTE

University of Minnesota

The concept of the criterion in much applied research has implied the possibility of identifying a single, ultimate measure against which predictors should be correlated. It is argued that the criterion has been overemphasized with the result that complexities of predicting the many facets of job success have been ignored in favor of overly simplified studies designed to relate predictors to single measures of job success. Applied psychologists should give more emphasis to construct validation and make an effort to learn more about the meaning of test scores and other predictors in terms of multiple dimensions of behavior. Information available on the Engineering Research Key of the Strong Vocational Interest Blank is presented in order to illustrate the pattern of validation research recommended.

Over the years, our concept of *the* criterion has suggested the existence of some single, all encompassing measure of job success against which other measures (predictors) might be compared. Consider, for example, a few quotations from widely used texts:

The ultimate criterion is the complete final goal of a particular type of selection or training (Thorndike, 1949, p. 121).

A criterion may be defined as a standard which can be used as a yardstick for measuring employees' success or failure (Stone & Kendall, 1956, p. 271).

A criterion is a standard. It is an index against which other indexes may be compared and evaluated. In our context, a criterion is a measure of job success. It therefore defines the desired end product of selection (Krug, 1961, p. 107).

For individuals on any given job there is, theoretically, some "ultimate" or "true" job standard by which the individuals might be evaluated (Tiffin & McCormick, 1958, p. 35).

And, consider the word itself. What does the new dictionary say in defining it?

Criterion: 1, a characterizing mark or trait; 2, a standard on which a decision or judgment may be based; a standard of reference; *yardstick* (Gove, 1961).

These statements and definitions imply the desirability of discovering, developing, or at least gaining agreement upon a *single* measurement of job success.

On the other hand, most authors also go on to comment upon the obvious reality that there usually are many criteria instead of only one. For example, Ghiselli and Brown

(1955) very nearly fail (wisely, I believe) to mention the word; and, instead, they discuss *various* ways of measuring job proficiency. Thorndike (1949) comments that, "A really complete ultimate criterion is multiple and complex in almost every case," and Krug (1961) points out that "Clearly a program of personnel selection can be no better than the criteria which define it."

Even so, most authors and certainly most research investigators act as if they were trying to achieve the pleasant state of having *one* measure with which to define job success. Thus, the psychological literature is full of warnings not to use a *deficient* or a *contaminated* criterion (Ghiselli & Brown, 1955; Krug, 1961; Nagle, 1953; Thorndike, 1949; Tiffin & McCormick, 1958) and we are warned often of the dangers of bias and also of the possibility of multiple dimensions (both static and dynamic) in criterion measures (Ghiselli, 1956; Ghiselli & Haire, 1960). In addition, a good deal has been written about the appropriate weighting of different measures of job success to form a single composite (Ghiselli & Brown, 1955; Krug, 1961; Tiffin & McCormick, 1958). It is in this area that the dilemma becomes even more clear-cut. If various measures correlate highly, an investigator gains some confidence that he is measuring the "core" of job success; yet, the necessity of their being weighted to form a single composite becomes less. Conversely, if the measures show low correlations a researcher may be gratified that he is tapping rather independent dimensions

of job success, or he may feel discontented because of the apparent lack of unity in this job success construct. As Thorndike (1949) states:

In general, high correlation between different intermediate criterion measures strengthens the rational basis for accepting any one of them as a useful criterion, since each of them receives some support from the rational justification of the other. Lack of correlation weakens faith in one or both measures, *except in so far as each measures distinct aspects of performance for which there is no rational basis to expect intercorrelation* [italics added] (p. 124).

In the latter instance, the question of whether to weight or not to weight becomes difficult and, unfortunately, the decision is usually in favor of weighting the uncorrelated measures in an ill-advised effort to develop a sort of "distilled essence" measure of total job success. It would seem wiser to investigate the separate relationships between each of the predictors and each of the available measures of job success. The results would lend increased meaning to the qualities measured by the tests *and* to the behaviors encompassed by the various aspects of job success.

The point of all this is to suggest that much selection and validation research has gone astray because of an overzealous worshiping of *the* criterion with an accompanying will-o-the-wisp searching for a best single measure of job success. The result has been an oversimplification of the complexities involved in test validation and the prediction of employee success. Investigators have been reluctant to consider the many facets of success and the concomitant investigation of the prediction of many success measures and instead persist in an unfruitful effort to predict *the* criterion. Thus, I say: junk *the* criterion! Let us cease searching for single or composite measures of job success and proceed to undertake research which accepts the world of success dimensionality as it really exists. Indeed, Thorndike (1960) concluded on the basis of a postmortem analysis of his 10,000-career study that, "Success is a many-faceted thing, and we need to relate the predictors to the different facets of success. This is a complex task and one not easy of fulfillment" (p. 16).

Our emphasis in test validation should change from the narrow point of view dictated by Predictive and Concurrent Validation and give increased emphasis to learning the *meaning* (Ebel, 1961) of test scores in terms of multiple dimensions of employee behavior. What I am suggesting is not new. My argument is simply an extension of previous statements outlining the methods of construct validation (Campbell, 1960; Chronbach & Meehl, 1955). However, I do believe that the rationale for learning more about the total constructs measured by tests has been largely ignored by industrial and other applied psychologists, who, for the most part, have been greatly concerned with *practical* validity to the near exclusion of learning what test scores mean in terms of the *many* behavioral measures with which they may be compared. As Campbell (1960) points out, institutional decisions dependent on determining practical validities have usually taken *the* criterion as an immutable given, even though such practice usually results in a gross oversimplification and a net loss of scientific information.

I am proposing, therefore, that coefficients of *practical* validity (concurrent and predictive) be accorded a lower position in the status hierarchy and that they be used simply as one of a number of kinds of evidence to lend meaning to test behavior.

To illustrate the pattern of validation research which I am recommending, let us examine the information available on the Engineering Research Key of the Strong Vocational Interest Blank (SVIB). The Research Key was developed (Dunnette, 1957) by comparing SVIB responses of research engineers with those made by "engineers in general." The following lines of evidence bearing on this key have now accumulated:

1. In a cross-validation group, research engineers scored significantly higher on the key than did development, production, or sales engineers (Dunnette, 1957).

2. The Research Key has positive scoring weights on the following SVIB scales: Artist, Psychologist, Architect, Physician, Dentist, Mathematician, Physicist, Chemist, Minister, and Author-Journalist (Dunnette, 1957).

3. Data available on 66 engineers undergoing assessment at the University of California's Institute for Personality Assessment and Research showed significant ($p < .05$) positive correlations between Research Key scores and scores on the Minnesota Engineering Analogies Test and on the Flexibility and Achievement (via independence) scores of the California Psychological Inventory (CPI). Significant negative correlations were obtained with the Dominance, Sociability, and Self-Assurance scores on CPI and with Economic and Political scores on the Allport-Vernon-Lindzey Scale of Values.¹

4. Data available on 35 research engineers working in a small Twin Cities electronics manufacturing firm showed significant ($p < .05$) positive correlations between Research Key scores and scores on the Miller Analogies Test, the Wesman Personnel Classification Test, and on the Intellectual Efficiency, Achievement (via independence), and Flexibility keys of CPI.

5. One hundred thirty-six engineers employed with the Minnesota Mining and Manufacturing Company (3M) completed a lengthy personal and biographical questionnaire. Engineers with high scores on the Research Key say that their most important goal in the next 5 years is to develop a new idea or invention as opposed to becoming an executive or earning a large amount of money. They also say more often than low scorers that they prefer working alone or with a small group instead of in jobs involving much interaction with other people. Additionally, they are less confident that they would make effective supervisors and they claim, as boys or young men, to have been introverted, *not* self-confident, and to have been frequently involved in making repairs around the house.

6. Supervisory ratings of various aspects of job performance were available on 72 3M engineers who had been on the job an average of 18 months. Scores on the Research Key correlated $+.32$ with ratings of Technical Competence and $-.19$ with ratings of Human Relations Skills for 29 research engineers. For a group of 19 sales engineers, the cor-

responding correlations were $-.14$ and $-.48$, respectively.

7. Of the 136 engineers described in Number 5 above, those with high scores on the Research Scale more often (than low scorers) used the following adjectives in describing themselves: resourceful, inventive, curious, mechanically inclined, and ingenious. Conversely, low scorers more often used the following adjectives: polished, poised, efficient, persuasive, and hardworking.

It is apparent from the foregoing that the search for the *meaning* of scores on the Research Key has made use of many types of measures ranging from supervisory ratings of technical competence and human relations skills to adjectival self-descriptions, other test scores, and biographical information. The knowledge accumulated from this series of studies tells us a good deal about the Research Key and its potential usefulness in counseling and/or job placement. In contrast, any series of studies designed solely to predict institutional criteria (e.g., ratings of job success) would have yielded less useful knowledge about the meaning of scores on the Research Key.

I suggest, therefore, that we cease talking about *the* criterion problem and that the notion of an ultimate criterion be disregarded. In so doing, a change in research emphasis should occur which will focus on defining the meaning of scores on any given test in terms of a variety of other (both test and nontest) behaviors rather than only in terms of some complex single or composite measure of job success. As a consequence, validation studies will become less restrictive and simple-minded, and will result in broader knowledge about the full meanings and uses of our predictor batteries.

REFERENCES

- CAMPBELL, D. T. Recommendations for APA test standards regarding construct, trait, or discriminant validity. *Amer. Psychologist*, 1960, **15**, 546-553.
- CRONBACH, L. J., & MEEHL, P. E. Construct validity in psychological tests. *Psychol. Bull.*, 1955, 281-302.
- DUNNETTE, M. D. Vocational interest differences among engineers employed in different functions. *J. appl. Psychol.*, 1957, **41**, 273-278.

¹ H. Gough, personal communication, June 1962.

- EBEL, R. L. Must all tests be valid? *Amer. Psychologist*, 1961, **16**, 640-647.
- GHISELLI, E. E. Dimensional problems of criteria. *J. appl. Psychol.*, 1956, **40**, 1-4.
- GHISELLI, E. E., & BROWN, C. W. *Personnel and industrial psychology*. New York: McGraw-Hill, 1955. P. 492.
- GHISELLI, E. E., & HAIRE, M. The validation of selection tests in the light of the dynamic character of criteria. *Personnel Psychol.*, 1960, **13**, 225-231.
- GOVE, P. B. (Ed.) *Webster's third new international dictionary*. Springfield, Mass.: G & C Merriam, 1961. P. 2662.
- KRUG, R. E. Personnel selection. In B. von H. Gilmer (Ed.), *Industrial psychology*. New York: McGraw-Hill, 1961. Ch. 6.
- NAGLE, B. F. Criterion development. *Personnel Psychol.*, 1953, **6**, 271-289.
- STONE, C. H., & KENDALL, W. E. *Effective personnel selection procedures*. Englewood Cliffs, N. J.: Prentice-Hall, 1956. P. 433.
- THORNDIKE, R. L. *Personnel selection*. New York: Wiley, 1949. P. 358.
- THORNDIKE, R. L. Ten thousand careers and criteria. Paper presented at American Psychological Association, Chicago, September 1960.
- TIFFIN, J., & MCCORMICK, E. J. *Industrial psychology*. Englewood Cliffs, N. J.: Prentice-Hall, 1958. P. 584.

(Received June 30, 1962)

THE RELATIONSHIP BETWEEN SOCIAL DESIRABILITY AND INTERNAL CONSISTENCY OF PERSONALITY SCALES¹

ALLEN L. EDWARDS, JAMES A. WALSH, AND CAROL J. DIERS

University of Washington

3 hypotheses concerning the relationships between psychometric characteristics of 61 personality scales were tested. A measure of internal consistency (Kuder-Richardson Formula 21) was found to be positively correlated (.62) with the degree of imbalance in the social desirability keying of the scales. Internal consistency was also negatively correlated ($-.46$) with the proportion of neutral items in the scales. The mean probability of a keyed response to the items in a scale was positively correlated (.83) with the proportion of items keyed for socially desirable responses. These results are consistent with predictions based upon social desirability considerations.

It has been shown by Edwards (1953) that the probability of a True response to a personality item is a linear increasing monotonic function of the social desirability scale value of the item. This finding has been confirmed in a number of other studies (Cowen & Tongas, 1959; Edwards, 1957a, 1959; Hanley, 1956; Hillmer, 1958; Kenny, 1956; Taylor, 1959; Wiggins & Rumrill, 1959; Wright, 1957). Further consideration of the relationship between the probability of a True response and social desirability scale value resulted in the concept of a *socially desirable response* to a personality item (Edwards, 1957b). A socially desirable response is defined as a True response to an item with a socially desirable scale value or as a False response to an item with a socially undesirable scale value. Because the relationship between probability of a True response and social desirability scale value is linear, the relationship between the probability of a socially desirable response and social desirability scale value is V shaped, with items at both extremes of the social desirability continuum having a high probability of eliciting a socially desirable response. As the social desirability scale values of items become more neutral, the probability of a socially desirable response decreases until, for items with precisely neutral scale values, the concept of a socially desirable response is undefined.

A Social Desirability (*SD*) scale was developed by Edwards (1957b) to measure the tendency to give socially desirable responses. The *SD* scale consists of 39 items of which 9 have socially desirable scale values and are keyed True and the remaining 30 have socially undesirable scale values and are keyed False, in accordance with the definition of a socially desirable response.

The tendency to give socially desirable responses in self-description is regarded by Edwards as a general trait which is elicited by all personality items regardless of the particular scale in which they happen to be contained. Thus, each item in a personality scale is regarded as having a specified probability, depending upon the scale value of the item, of eliciting a socially desirable response.

As one index of the degree to which scores on a personality scale may be influenced by social desirability tendencies, Edwards has used the proportion of items in the scale which are keyed for socially desirable responses. It has been demonstrated in a number of studies (Edwards, 1957b, 1961; Edwards, Heathers, & Fordyce, 1960; Edwards, Diers, & Walker, 1962) that this index is highly correlated with the correlation of a personality scale with the *SD* scale. Scales which have a large proportion of items keyed for socially desirable responses tend to have high positive correlations with the *SD* scale and scales which have a small proportion of items keyed for socially desirable responses, that is, a large proportion keyed for

¹This research was supported in part by Research Grant M-4075 from the National Institute of Mental Health, United States Public Health Service.

socially undesirable responses, tend to have high negative correlations with the *SD* scale.

If all of the items in a scale are keyed for responses in such a way that the keyed responses are believed to measure a common personality trait, then the interitem correlations of the keyed responses should be positive. Consider a scale in which all of the items are keyed not only for the trait but also for socially desirable responses. In a scale of this kind the tendency to give the trait response is confounded with the tendency to give the socially desirable response and we should expect to find not only that the interitem correlations are positive but also that scores on the scale are positively correlated with scores on the *SD* scale. If the trait keying is the same as the social undesirability keying, then the tendency to give the trait response is confounded with the tendency to give socially undesirable responses and we should expect to find that the interitem correlations are positive and that scores on the scale are negatively correlated with scores on the *SD* scale. Suppose, however, that the trait keying for some of the items is the same as the social desirability keying and the trait keying for the remaining items is the same as the social undesirability keying. If the trait itself is the only important determiner of responses to the items, then the interitem correlations should still be positive. To the degree to which the tendency to give socially desirable responses is operating, we should expect to find that the interitem correlations in the set of items keyed for socially desirable responses are positive. Similarly, the interitem correlations in the set of items keyed for socially undesirable responses should also be positive. However, the correlations between those items in the set keyed for socially desirable responses and those in the set keyed for socially undesirable responses should be decreased, if the tendency to give socially desirable responses is operating. When there is a balance in the social desirability keying of a scale, we should also expect to find that scores on the scale have relatively low correlations with the *SD* scale.

The Kuder-Richardson Formula 21 (K-R 21) is generally regarded as an index of the

degree to which the interitem correlations in a scale are positive and, therefore, a measure of the degree to which the items in a scale are being responded to in terms of a common trait. If responses to the items in a personality scale are influenced not only by the trait which the scale was designed to measure but also by the tendency to give socially desirable responses, then we should expect to find that K-R 21 values are related to the nature of the social desirability keying of the items. Trait scales in which the items are consistently keyed for either socially desirable or socially undesirable responses should have higher K-R 21 values than scales which are more balanced in their social desirability keying. This is one hypothesis to be tested in the present study.

It has been pointed out that the probability of a socially desirable response decreases as the scale values of items become more neutral. If responses to the items in a scale are primarily determined by the trait which the scale is supposedly measuring, then the presence of a large number of neutral items in the scale should have little or no influence upon the interitem correlations. If anything, since the tendency to give socially desirable responses is minimal for neutral items, we might expect trait responses to be more dominant to these items. If this hypothesis is correct, then the proportion of neutral items in a scale should be positively correlated with K-R 21 values. In other words, if all of the items in a scale consist of items with neutral scale values, then we might expect all responses to be trait determined and uninfluenced by social desirability tendencies. On the other hand, if social desirability tendencies are important determiners of responses to personality items, then the presence of a large number of neutral items in a scale must be fairly confusing for subjects with strong tendencies to give socially desirable responses. Some of these subjects may regard the True response to a neutral item as more socially desirable and others the False response as more socially desirable, and there should be little or no tendency for such subjects to respond consistently to a set of neutral items. If their responses to neutral items are more or less random, then the presence of a large number

of neutral items in a scale should result in a lower K-R 21 value for the scale than for a scale which has a small number of neutral items. A second hypothesis to be tested, therefore, is that K-R 21 values for a set of personality scales are negatively correlated with the proportion of items with neutral scale values in the personality scales.

Each item in a personality scale is keyed for a response which is believed to indicate the presence of a given personality trait. It has been shown that the probability of a True response to a personality item is a function of the social desirability scale of an item. Since the keyed response to an item in a True-False personality scale is either True or False, the probability of a keyed response may also be regarded as a function of the social desirability scale value of the item. If an item has a high social desirability scale value and is also keyed True, then the probability that the item will elicit the keyed response should be greater than it would be if the item were keyed False. As an average index of the probability of a keyed response being made to the items in a personality scale, we define

$$P(K) = \bar{X}/n$$

where $P(K)$ is the mean probability of a keyed response, \bar{X} is the mean score on the scale, and n is the number of items in the scale. If the tendency to give socially desirable responses is influencing responses to the items in a scale, then we should expect $P(K)$ to be greater for scales which contain a large proportion of items keyed for socially desirable responses than for scales which contain a small proportion of items keyed for socially desirable responses. If we let $P(SD)$ indicate the proportion of items keyed for socially desirable responses, then $P(SD)$ should be positively correlated with $P(K)$. This is the third hypothesis to be tested in the present study.

METHOD

Scores on 58 Minnesota Multiphasic Personality Inventory (MMPI) scales and 3 non-MMPI scales were available from another study (Edwards, Diers, & Walker, 1962). For each scale we obtained a K-R 21 value, based upon the mean, variance, and num-

ber of items in the scale. For each scale, we also obtained $P(K)$, the mean probability of a keyed response. The proportion of neutral items, $P(N)$, in each MMPI scale was based upon the number of items having scale values between 2.5 and 3.5 on a 5-point social desirability continuum. Social desirability scale values for the MMPI items were determined by Heineman and are published in Dahlstrom and Welsh (1960). Similarly, $P(SD)$, the proportion of items keyed for socially desirable responses, was based upon the Heineman scale values, an item being considered as keyed for a socially desirable response if it had a scale value above the neutral point and if it was also keyed True or if it had a scale value below the neutral point and was keyed False.

According to the social desirability hypothesis, scales which have either a large or a small value of $P(SD)$ should have higher K-R 21 values than scales which contain a balance in the social desirability keying. Thus, in relating K-R 21 values to $P(SD)$, a value of $P(SD) = .10$ may be regarded as equivalent to a value of $P(SD) = .90$. In order to correlate K-R 21 values with the imbalance in the social desirability keying of the scales, we used as a measure of the degree of imbalance: $I-SD = |P(SD) - .50|$. Thus, when $I-SD = .00$, it means that an equal number of items was keyed for socially desirable and socially undesirable responses and when $I-SD = .50$, it means that either all of the items were keyed for socially desirable responses or that all of the items were keyed for socially undesirable responses.

RESULTS AND DISCUSSION

The product-moment correlation between K-R 21 values for the 61 scales and the measure of imbalance, $I-SD$, is given in Table 1 and is .62. This correlation is consistent with the hypothesis that, in general, K-R 21 values increase as the degree of imbalance in the social desirability keying increases. In other words, if the items in a scale are consistently keyed for either socially desirable or socially undesirable responses, then KR-21 values are higher than they are for scales which contain a more balanced social desirability keying. This result is of some practical significance, for it has been shown that if a scale is consistently keyed for socially desirable responses, then it will also tend to have a high positive correlation with the SD scale. On the other hand, if a scale is consistently keyed for socially undesirable responses, then it will tend to have a high negative correlation with the SD scale. It has been suggested that if one wishes to minimize the correlation between a given scale and the

TABLE 1

CORRELATIONS BETWEEN PSYCHOMETRIC CHARACTERISTICS OF 61 PERSONALITY SCALES AND MEANS AND STANDARD DEVIATIONS OF THE CHARACTERISTICS

Variables						
1	2	r_{12}	\bar{X}_1	\bar{X}_2	s_1	s_2
K-R 21	$I-SD$.62	.57	.83	.20	.14
K-R 21	$P(N)$	-.46	.57	.24	.20	.17
$P(SD)$	$P(K)$.83	.43	.43	.36	.18

SD scale, then this might be achieved by developing a scale in which there is a balance in the social desirability keying. But the correlation between K-R 21 and *I-SD* indicates that if this is done, the scale will, in general, have a lowered K-R 21 value. In minimizing the influence of social desirability, in this way, we may be sacrificing homogeneity, as measured by K-R 21, and thus have a less pure measure of the trait we would like to measure.

It has also been suggested that social desirability influences might be minimized by developing scales which contain a large proportion of neutral items. But, as the negative correlation of $-.46$ between $P(N)$ and K-R 21 indicates, as the proportion of neutral items in a scale is increased, the result is a less homogeneous test.

The correlation between the mean probability of a keyed response, $P(K)$, and the proportion of items keyed for socially desirable responses, $P(SD)$, is $.83$. This correlation is consistent with the hypothesis that the mean probability of a keyed response to items in a personality scale is a function of the social desirability keying of the items. Scales, which have a large proportion of items keyed for both a trait response and for a socially desirable response, have a greater probability of eliciting keyed responses than scales in which the trait responses tend to be keyed for socially undesirable responses.

It has been suggested by Fricke (1957), Jackson and Messick (1958), and Couch and Keniston (1960), that another important determiner of responses to MMPI items is acquiescence or, as Couch and Keniston state, the tendency to respond True to an item

regardless of the content of the item. If acquiescence is an important determiner of responses to MMPI items, then the interitem correlations of scales in which all of the items are keyed True or in which all of the items are keyed False should, in general, be positive and greater than in the case of scales which have some degree of balance in their True keying. Thus, according to the acquiescence hypothesis, scales which contain either a large or a small portion of items keyed True should have higher K-R 21 values than scales which are better balanced in their True keying. To test this hypothesis, we obtained *I-T*, the measure of imbalance in the True keying, for each of the scales involved in this study. The product-moment correlation between *I-T* and K-R 21 values was $.08$ and the magnitude of the correlation offers little support for the acquiescence hypothesis.

The acquiescence hypothesis would also predict that $P(K)$, the mean probability of a keyed response, should be related to $P(T)$, the proportion of items keyed for True responses in a scale. Whether this relationship should be positive or negative depends upon whether one considers acquiescence to be more, or less, dominant, on the average, than dissentience in a given group of subjects.

In the present study, the correlation between $P(T)$ and $P(K)$ was found to be $-.45$ and the correlation might be considered as offering evidence in favor of the hypothesis that the probability of a keyed response is dependent upon the proportion of items keyed for True responses. However, it has been pointed out by both Hanley (1961) and Edwards (1961) that $P(T)$ is often confounded with $P(SD)$ and this is particularly the case with MMPI scales. In general, if an MMPI scale has a large proportion of items keyed True, then it also tends to have a large proportion of items keyed for socially undesirable responses. For the scales involved in the present study, the correlation between $P(T)$ and $P(SD)$ is the same value, $-.45$, as the correlation between $P(T)$ and $P(K)$. We suggest, therefore, that the correlation between $P(T)$ and $P(K)$ is at least in part the result of the confounding between

the True keying and the social desirability keying of the scales.

REFERENCES

- COUCH, A., & KENISTON, K. Yeasayers and nay-sayers: Agreeing response set as a personality variable. *J. abnorm. soc. Psychol.*, 1960, **60**, 151-174.
- COWEN, E. L., & TONGAS, P. The social desirability of trait descriptive terms: Applications to a self-concept inventory. *J. consult. Psychol.*, 1959, **23**, 361-365.
- DAHLSTROM, W. G., & WELSH, G. S. *An MMPI handbook*. Minneapolis: Univer. Minnesota Press, 1960.
- EDWARDS, A. L. The relationship between the judged desirability of a trait and the probability that the trait will be endorsed. *J. appl. Psychol.*, 1953, **37**, 90-93.
- EDWARDS, A. L. Social desirability and probability of endorsement of items in the Interpersonal Check List. *J. abnorm. soc. Psychol.*, 1957, **55**, 394-395. (a)
- EDWARDS, A. L. *The social desirability variable in personality assessment and research*. New York: Dryden, 1957. (b)
- EDWARDS, A. L. Social desirability and the description of others. *J. abnorm. soc. Psychol.*, 1959, **59**, 434-436.
- EDWARDS, A. L. Social desirability or acquiescence in the MMPI? A case study with the *SD* scale. *J. abnorm. soc. Psychol.*, 1961, **63**, 351-359.
- EDWARDS, A. L., HEATHERS, LOUISE B., & FORDYCE, W. E. Correlations of new MMPI scales with Edwards *SD* scale. *J. clin. Psychol.*, 1960, **16**, 26-29.
- EDWARDS, A. L., DIERS, CAROL J., & WALKER, J. N. Response sets and factor loadings on sixty-one personality scales. *J. appl. Psychol.*, 1962, **46**, 220-225.
- FRICKE, B. G. A response bias scale for the MMPI. *J. counsel. Psychol.*, 1957, **4**, 149-153.
- HANLEY, C. Social desirability and responses to items from three MMPI scales: *D*, *Sc*, and *K*. *J. appl. Psychol.*, 1956, **40**, 324-328.
- HANLEY, C. Social desirability and response bias in the MMPI. *J. consult. Psychol.*, 1961, **25**, 13-20.
- HILLMER, M. L., JR. Social desirability in a two-choice personality scale. Unpublished master's thesis, University of Washington, 1958.
- JACKSON, D. N., & MESSICK, S. Content and style in personality assessment. *Psychol. Bull.*, 1958, **55**, 243-252.
- KENNY, D. T. The influence of social desirability on discrepancy measures between real self and ideal self. *J. consult. Psychol.*, 1956, **20**, 315-318.
- TAYLOR, J. B. Social desirability and MMPI performance: The individual case. *J. consult. Psychol.*, 1959, **23**, 514-517.
- WIGGINS, J. S., & RUMRILL, C. Social desirability in the MMPI and Welsh's Factor Scales A and R. *J. consult. Psychol.*, 1959, **23**, 100-106.
- WRIGHT, C. E. Relations between normative and ipsative measures of personality. Unpublished doctoral dissertation, University of Washington, 1957.

(Received July 30, 1962)

EFFECT OF VARIATION IN TASK COMPLEXITY AND DISPLAYED INFORMATION ON OPERATOR PERFORMANCE¹

KENNETH ZIEDMAN AND JOHN LYMAN

University of California, Los Angeles

An experiment was conducted to determine the effects of amount of information displayed, level of task difficulty, and practice on performance speed and accuracy for a visual motor task. Based on results from 16 Ss, it was concluded that an increase in the number of cues available for visual reference will not necessarily affect performance. Further corroborative evidence was obtained for the hypothesis that the criterion of redundancy in displayed information should be based on perceptual usefulness of the cues provided.

The specification of optimum visual displays has frequently made use of concepts which refer to the information-carrying attributes of display configurations. For example, such terms as clean display, extraneous or irrelevant information, clutter, and redundant information are common in the literature. These concepts have evolved from both perceptual psychology and from the more recent attempts to apply information theory to perception (Broadbent, 1958; Cherry, 1957). The present study is an attempt to obtain additional information on the concept of redundancy as it applies to visual displays and to help distinguish the broader context of the term from conditions under which it can be given a rigorous mathematical definition.

Although redundancy has been given an explicit definition in information theory (Shannon & Weaver, 1949), it has not been generally possible to isolate those aspects of a display-control system which are amenable to an information theory analysis. The information theory concept of redundancy, therefore, cannot be directly applied to a description of visual display configurations. One reason for this is that the transmission properties of the human operator cannot be specified with sufficient quantitative precision

to allow a choice of the proper information measure to be applied to his input. Another reason is that information theory deals with statistical processes, whereas the significant spatial properties of display configurations are not necessarily statistical in nature. The particular properties of the display may override assumed statistical properties of the information input, leading to seemingly contradictory experimental results because of the experimenter's inability to predict the relevance to the subject of the variables he introduces to the situation.

For example, in an experiment relating redundancy in randomly constructed figures to sorting time (Rappaport, 1957), sorting time was found to increase with increased redundancy. Redundancy was produced in the experiment by repeating the figure while holding area constant. In the second part of the experiment, sorting time was found to increase with decreased detail size, thus implying that the results in the first half may have been due to decreased detail size rather than increased redundancy.

The difficulty in actually defining the pertinent parameters seems to indicate an obvious need for delineation of the perceptual usefulness² of different types of cues for the performance of visual-motor tasks. The purpose of the present investigation was to examine the interaction of the amount of displayed information in a visual display with

¹ This investigation was carried out in the Biotechnical Laboratory at the University of California at Los Angeles and was supported by the Office of Naval Research, Engineering Psychology Branch, under Contract Nonr 233(49). The opinions expressed are those of the authors and do not necessarily reflect the views of the contracting agency.

² "Perceptual usefulness" has been defined as the rank order of utility of particular display cues as aids to performance (see Groth & Lyman, 1961, p. 86).

task difficulty and amount of practice for a self-paced task as part of a series of experiments on the importance of redundant information at different stages of practice (Groth & Lyman, 1961; Ziedman & Lyman, 1960).

It is established that the relative usefulness of displayed information interacts with task difficulty. For example, performance is about equal on compensatory and pursuit displays for low-frequency inputs, whereas the pursuit display is superior for high-frequency inputs (Bowen & Chernikoff, 1957). Furthermore, both amount of information and task difficulty may interact with practice (Hodge, 1959). The results obtained with differences in practice may depend upon the storing of information in memory, which can then be considered to be redundant to the displayed information.

The task of the operator in the present experiment can be represented as a simple decision-making problem in which he had to identify the spatial location of the stimulus among a set of alternatives. Four different display-cue structures were used to provide increasing amounts of spatial location information.

METHOD

Apparatus

The display-cue analyzer used in the experiment was described in an earlier report (Ziedman, 1960).

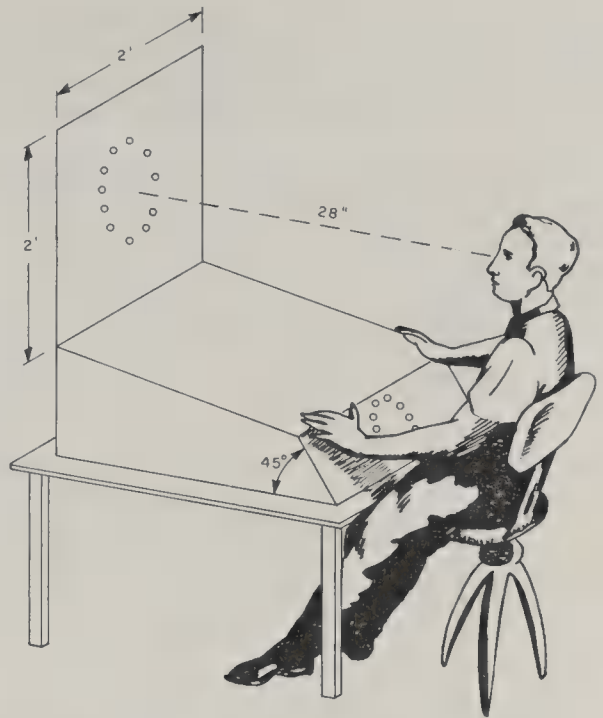


FIG. 1. Display-control panel.

The arrangement of the display and response panels is shown in Figure 1. The display panel consisted of a number of small lamps mounted behind a 24 × 24 inch sheet of translucent white cardboard. When a lamp was illuminated, a ¼-inch spot of light could be seen through the cardboard panel. The desired display configuration was drawn on the face of the cardboard panel and could be changed for the different experimental conditions. The lamps were arranged in three circles of 12 lamps each. The diameter of the circles was 2 ¼ inches, and the center-to-center distance between them was 6 inches. The response panel con-

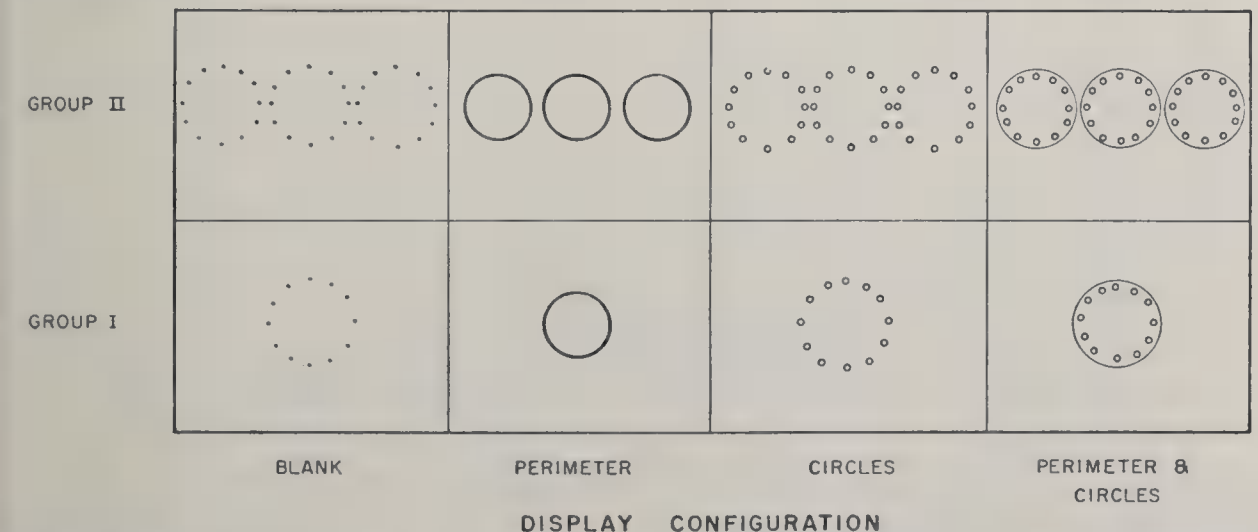


FIG. 2. Display configurations for Group I—one-dial group—and Group II—three-dial group. (Dotted points for the blank displays indicate lamp positions. Center-to-center distances between the groups of lamps are not to scale.)

tained 12 push-button switches arranged in a single 4-inch diameter circle. A force of 4 ounces was necessary to actuate the switches. The operator's eye distance from the display panel was about 28 inches.

Upon illumination of a lamp, the operator's task was to press the response switch in spatial correspondence to the lamp. This turned the lamp off, and the next lamp in a preprogramed sequence was turned on after a delay of about .15 second. If an incorrect switch was pressed, the lamp remained on.

Independent Variables

Two levels of task difficulty were paired with four display configurations, as shown in Figure 2, in a Lindquist Type VI design (Lindquist, 1953). Group I received signals from only the center circle of 12 lamps. Group II received signals from all three circles of lamps. A single lamp in the same position within the circles was illuminated for both groups on each trial. For Group I the lamp was chosen from only the center circle, whereas for Group II the lamp was chosen from any of the

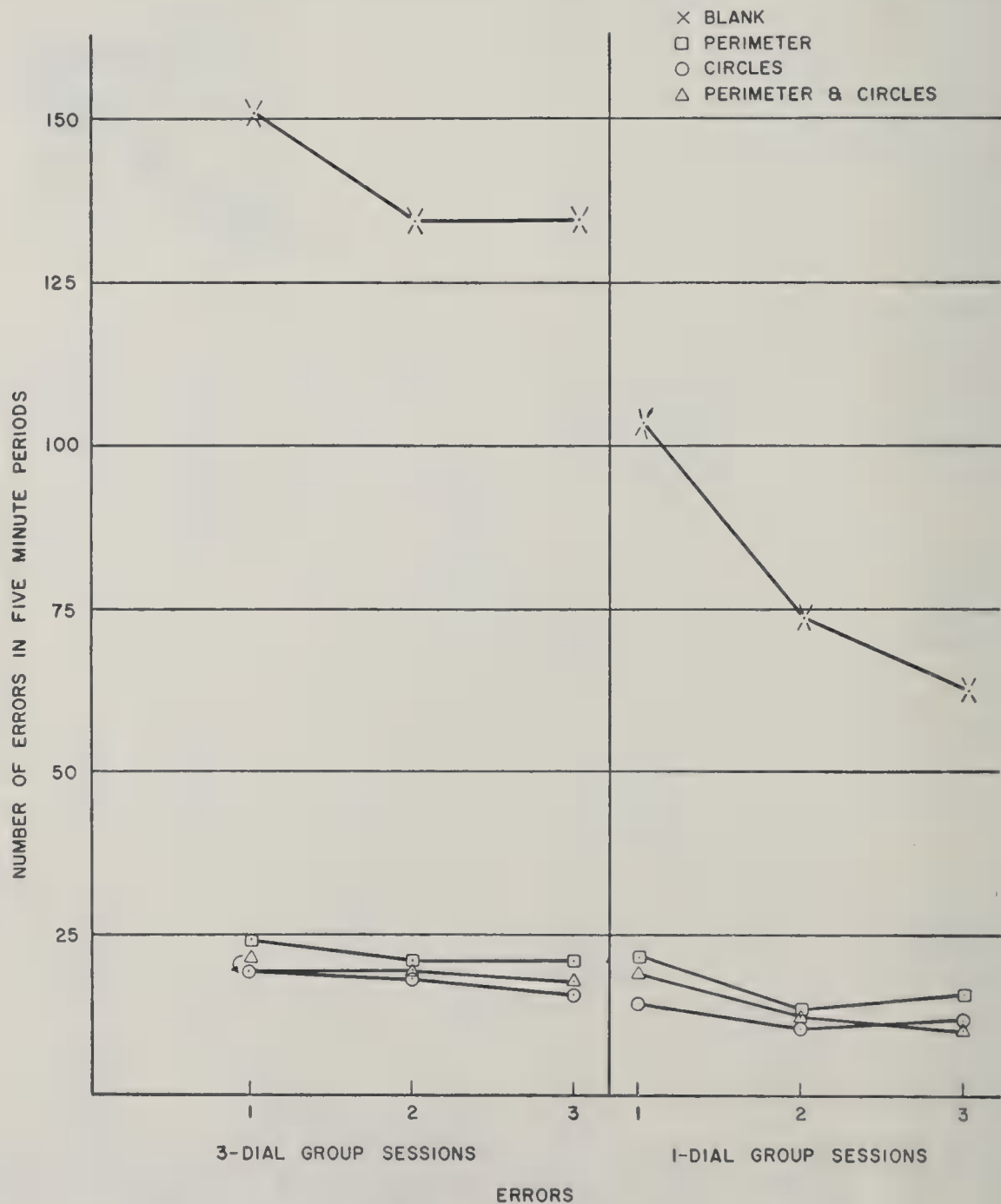


FIG. 3. Number of errors in 5-minute periods averaged over subjects.

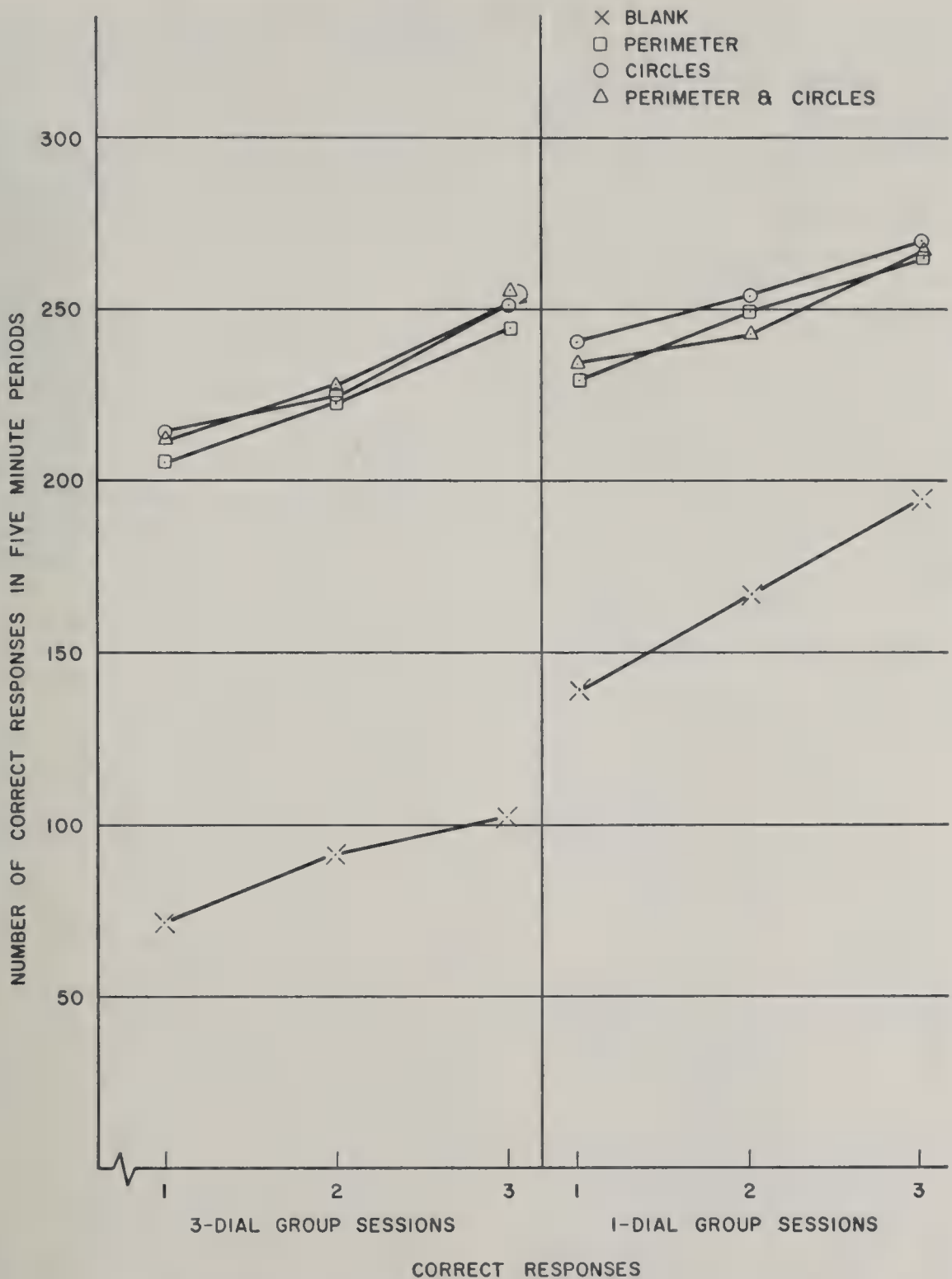


FIG. 4. Number of correct responses in 5-minute periods averaged over subjects.

three circles. The same group of 12 response switches was used for both groups.

The four display configurations were:

1. Blank (B): No additional cues were provided on the display panel.

2. Perimeter (P): A $2\frac{3}{4}$ -inch diameter perimeter outlining the area containing the signal lamps was displayed.

3. Circles (C): Twelve $\frac{1}{4}$ -inch circles marking the location of the signal lamps were shown.

4. Perimeter and Circles (PC): The conditions of P and C were combined.

Group I had only the center circle marked for Conditions P, C, and PC, whereas Group II had all three circles marked. The lamps could not be seen through the panel so that the subjects in Group I were unaware of the other lamps.

Dependent Variables

The performance measures were: number of errors (accuracy), and total number of responses (speed).

Stimulus Program

The lamp sequence was randomized with the constraint that each lamp occurred an equal number of times in a basic sequence of 144 trials which was repeated as many times as necessary. For Group II the dial sequence was determined from a random number table.

Procedure

Sixteen undergraduate students, 10 males and 6 females, served as subjects. Eight subjects were randomly assigned to Groups I and II with the

males and females being equally divided between the two groups. Each subject served for three 1-hour sessions spaced 1 week apart. During a session, a subject worked for 8 minutes on each of the four displays in his group. His performance was recorded for the last 5 minutes of each 8-minute period to minimize the effects of "warm up" for the particular experimental condition. The display order was counterbalanced over subjects, but each subject received the same order for all three sessions.

Instructions

Standardized instructions were read to the subjects before each hour session. The instructions stressed both speed and accuracy.

RESULTS

General

The results of errors, correct responses, and total responses for each condition are given in Figures 3, 4, and 5, respectively. Each graph presents the results for both groups as a function of session with the display type as a parameter.

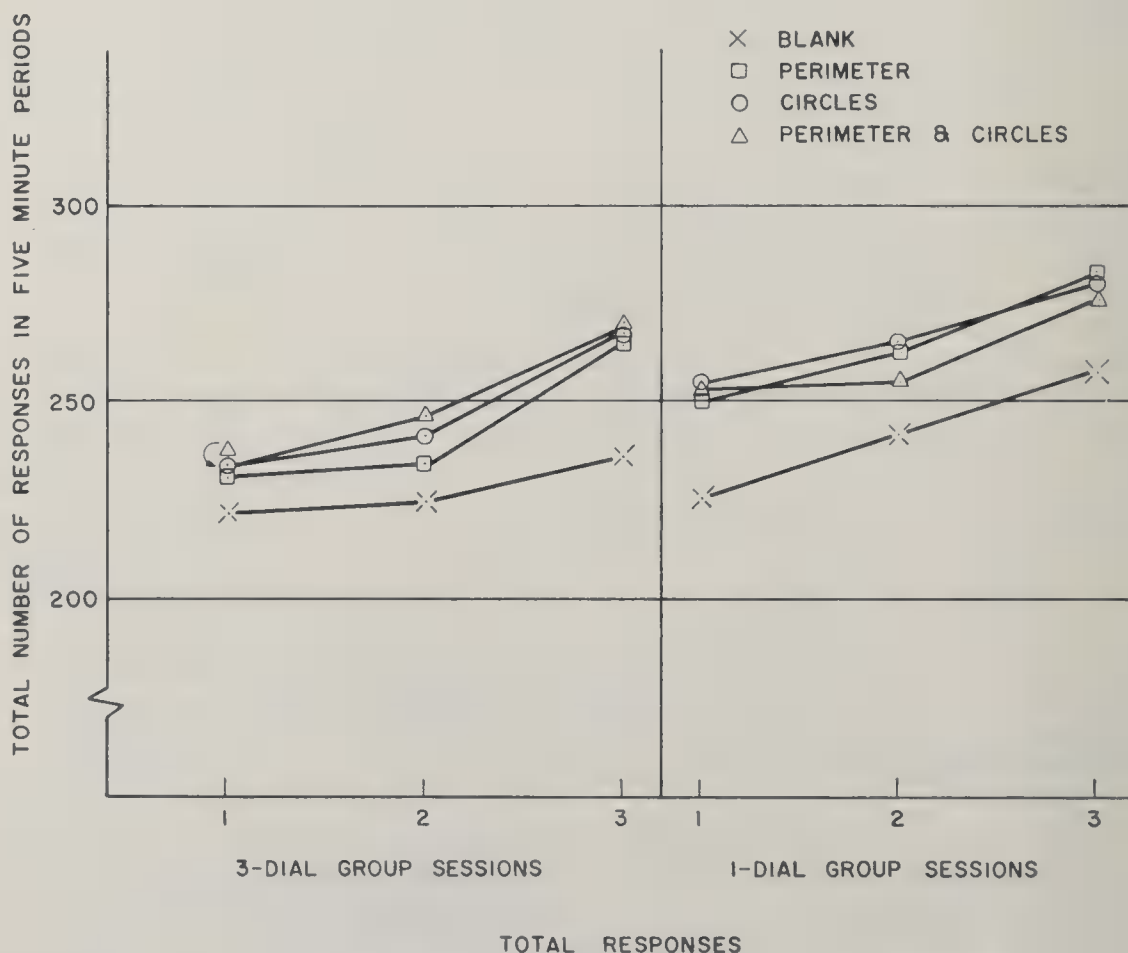


FIG. 5. Total number of responses in 5-minute periods averaged over subjects.

Inspection of the graphs indicates a marked superiority of Displays P, C, and PC over Display B. The superiority was maintained over all sessions and was most pronounced for the correct response and error measures. However, the results also indicated that there was no practical difference between Displays P, C, and PC, regardless of the session or measure.

Statistical Tests

Bartlett's test for homogeneity of variance revealed homogeneous variances for the correct response and total response measures ($\chi^2 = 30.77$ and 36.03 , $df = 23$, respectively) and a heterogeneous variance for the error measure ($\chi^2 = 199$, $df = 23$). The heterogeneous error variances were probably caused by the large difference in means between Display B and the other three displays. The significance of the differences between conditions were, therefore, only examined for the correct response and total response measures. The results are summarized in Table 1.

Both the display and session differences were significant for the total response and correct response measures. The total response measure did not distinguish between Groups I and II (complexity dimension). However, all three main effects and the Complexity \times Display interaction were significantly different as measured by correct responses.

Duncan's range test (Duncan, 1955) was used to determine the individual mean differences. For both measures Displays P, C, and PC were significantly different from Display B, but did not differ among themselves. In both cases all three session means were significantly different from each other.

Although the total response measure gave a significant difference between Display B and the other three displays, the absolute difference was of little practical importance. The difference between Display B and Displays P, C, and PC as given by correct responses was quite large. The major difference between conditions, therefore, was due to differences in errors and not due to differences in response rates. For example, the difference in total responses between Display B for the three-dial condition (the most difficult dis-

TABLE 1
ANALYSIS OF VARIANCE SUMMARY

Source	df	Total responses	Correct responses
Displays (D)	3/42	3.27*	184.26***
Complexity (C)	1/14	1.78	13.55**
Sessions (S)	2/28	20.5***	39.97***
D \times C	3/24	1.0	12.37***
C \times S	2/28	1.0	1.0
D \times S	6/84	1.0	1.0
D \times S \times C	6/84	1.0	1.48

* $p < .05$.
 ** $p < .01$.
 *** $p < .001$.

play condition) and Display PC for the one-dial condition (an easy condition) was only 4.1% for the last session. The difference in errors for the same condition was over 1,000%.

The Complexity \times Display interaction for correct responses indicated that the relative usefulness of the displays was different at different levels of complexity. As can be seen from Figure 4, the effect of increased complexity was a large reduction of correct responses for Display B, with only a small reduction for the other displays.

An increase in total responses and correct responses was observed over the three sessions. In addition, Figure 3 shows that the errors on Display B for the one-dial condition decreased over the three sessions, whereas the errors for Display B on the three-dial condition did not decrease after the second session.

DISCUSSION

Although the results of this study indicated an interaction of displayed information with task difficulty, the principal result was the failure of the criterion measure to differentiate among the three displays with spatial location cues (Displays P, C, and PC) and the significant inferiority of the blank display (Display B) in comparison to the other three. Two points seem to be implied by these results:

1. In the case of spatial location, increasing the number of cues available for a visual reference system will not necessarily result

in improvement of performance. A criterion of redundancy, as stated above, must be based on the usefulness of the displayed information to the operator for the given task.

2. The sharp distinction between the performance on Display B and the other three indicates that although the experimenter had intended to vary his stimuli along a single dimension of redundancy, he actually produced a more complex situation. The information content of a display, and by implication the amount of redundancy, is a function of many factors. Because these factors are not well understood, it seems difficult to construct an a priori set of displays varying along a single dimension of redundancy.

The different amounts of intended redundancy in Displays P, C, and PC were either not used by the subjects, or if they were used, they did not affect the performance measures.³ If it can be assumed that the subjects were indeed not sensitive to the additional information, then it cannot be called redundant with respect to the particular experimental situation used. The problem is not solved by using the definition of redundancy found in information theory because one has no guarantee that the human operator will be sensitive to inputs based on that definition. The characteristics of the task, the criterion of performance used by the operator, and the experience gained by the operator during practice must all be considered in addition to the visual cues internal

to the display. A particularly interesting aspect of the problem is the balance between the information displayed to the operator and that available to him through memory as related to task variables and practice. It is hoped that a continuation of these experiments will enable the determination of some of these relationships.

REFERENCES

- BOWEN, J. H., & CHERNIKOFF, R. The relationships between magnification and course frequency in compensatory aided tracking. *USN Res. Lab. Rep.*, 1957, No. 4913.
- BROADBENT, D. E. *Perception and communication*. London: Pergamon Press, 1958.
- CHERRY, C. *On human communication: A review, a survey, and a criticism*. New York: Wiley, 1957.
- DUNCAN, D. B. Multiple range and multiple *F* tests. *Biometrics*, 1955, 11, 1-42.
- GROTH, HILDE, & LYMAN, J. A hierarchy of "perceptual usefulness" of geometrical cues in an overlearned dial-reading task. *J. appl. Psychol.*, 1961, 45, 86-90.
- HODGE, M. H. The influence of irrelevant information upon complex visual discrimination. *J. exp. Psychol.*, 57, 1-5.
- LINDQUIST, E. F. *Design and analysis of experiments in psychology and education*. Boston: Houghton Mifflin, 1953.
- RAPPAPORT, M. The role of redundancy in the discrimination of visual forms. *J. exp. Psychol.*, 1957, 53, 3-10.
- SHANNON, C., & WEAVER, W. *The mathematical theory of communication*. Urbana: Univer. Illinois Press, 1949.
- ZIEDMAN, K. Development and specifications of the display-cue analyzer. Report No. 60-73, 1960, University of California, Los Angeles, Engineering Department.
- ZIEDMAN, K., & LYMAN, J. An assessment of probability distribution of signal occurrence in combination with relevant and irrelevant cues for masses practice and transfer of training. Report No. 60-75, 1960, University of California, Los Angeles, Engineering Department.

(Received July 30, 1962)

³ The possibility that the displays might have caused a difference in some other aspect of behavior cannot be discounted. For example, greater effort and closer attention to the task might have overcome a lack of information in the display.

JOB ATTITUDES IN MANAGEMENT:

III. PERCEIVED DEFICIENCIES IN NEED FULFILLMENT AS A FUNCTION OF LINE VERSUS STAFF TYPE OF JOB¹

LYMAN W. PORTER²

University of California, Berkeley

This study focused on differences in perceptions of the degree of fulfillment and importance of several types of psychological needs associated with line versus staff types of management jobs. Data were provided from a questionnaire, and the nationwide sample of respondents included 1802 managers from a wide variety of types of companies. Results showed: (a) Line managers perceived greater need fulfillment than staff managers, with the largest line-staff differences occurring in the Esteem and Self-Actualization need areas. (b) Line and staff managers did not differ on the importance they attached to each type of need, with the exception of Autonomy needs which staff managers considered more important. (c) Effects of the line-staff variable were smaller than effects of the variable of vertical level of position within the management hierarchy.

The two previous papers in this series on job attitudes in management dealt with perceived deficiencies in need fulfillment and perceived importance of needs as a function of differences in the vertical location of the management position (Porter, 1962, 1963). In contrast, the present study investigates these perceptions in relation to an aspect of *horizontal* differences among management positions, namely, line versus staff types of jobs. This horizontal dimension within the management structure of organizations has been generally ignored in any studies that have collected systematic research data on management. Although a large number of authors of books and articles on management have discussed, often in great detail, line-staff relationships, the amount of research data in this area is quite limited.

The traditional distinction between line and staff functions in organizations is that indi-

viduals in line jobs are supposed to be concerned with the main operations of the organization and function within the direct chain of command, while those in staff positions are concerned with auxiliary services that provide advice and assistance to the line and function outside of the direct chain of command. Recently, this sharp distinction between line and staff positions and functions within organizations has been questioned by a number of writers (e.g., Brown, 1954; Fisch, 1961; McGregor, 1960) who contend that the two types of positions should be more integrated into functional teams or groups of managers with common goals and objectives. Nevertheless, in most existing organizations the distinction remains between line jobs and staff jobs, more or less along the lines of the traditional conception of differences in the functions and responsibilities of the two types of management positions.

The purpose of the present study is to investigate the differences in perceived deficiencies in need fulfillment between individuals holding line positions and those holding staff jobs. In other words, do the differences in role, and perhaps status, along this horizontal dimension of organization structure affect attitudes towards satisfaction of different types of psychological needs that can be fulfilled in the work situation? Is the impact of the horizontal dimension, reflected

¹ This study was carried out as part of the research program of the Institute of Industrial Relations, University of California, Berkeley. It was started while the author was a Ford Foundation Faculty Research Fellow. The Institute of Social Sciences at the University of California and the American Management Association contributed to the support of the research assistance, and the Computer Center of the University provided facilities for data computations.

² The author is indebted to Mildred Henry, Larry Stewart, and Robert Andrews for assistance in tabulation of the data.

in the line-staff distinction, as great as that of the vertical dimension, reflected in the distinction among different levels of jobs within the organization hierarchy? This paper also briefly considers differences in the perception of the importance of various needs between individuals occupying line positions versus those in staff positions.

METHOD

Questionnaire instrument

Data for the present study were obtained by the use of a questionnaire described in detail in previous articles (Porter, 1961, 1962). Thirteen items in the questionnaire, all pertaining to a Maslow-type need hierarchy classification system, furnished the basic data. A sample item, as it appeared in the questionnaire, was as follows:

The feeling of *self-esteem* in my management position.

- (a) How much is there now?
(min) 1 2 3 4 5 6 7 (max)
- (b) How much should there be?
(min) 1 2 3 4 5 6 7 (max)
- (c) How important is this to me?
(min) 1 2 3 4 5 6 7 (max)

As is explained in the Results section, answers to Parts a and b of each item provided the data on perceived deficiencies in need fulfillment. Answers to Part c of each item were used to assess the importance of each type of need to the respondent.

Categories of Needs and Specific Items

The hierarchical categories of needs studied in this investigation are listed below, along with the specific items used to obtain information on each category. The items are here listed systematically according to their respective need categories, but were presented randomly in the questionnaire. The categorization system, which has been described in detail in a preceding paper (Porter, 1961), is based on Maslow's classification scheme (Maslow, 1954), although it is not identical with it. The need categories and their specific items follow:

I. Security needs

1. The *feeling of security* in my management position

II. Social needs

1. The *opportunity*, in my management position, to give help to other people
2. The *opportunity to develop close friendships* in my management position

III. Esteem needs

1. The *feeling of self-esteem* a person gets from being in my management position
2. The *prestige* of my management position *inside* the company (that is, the regard received from others *in* the company)

3. The *prestige* of my management position *outside* the company (that is, the regard received from others *not* in the company)

IV. Autonomy needs

1. The *authority* connected with my management position
2. The *opportunity for independent thought and action* in my management position
3. The *opportunity*, in my management position, for *participation in the setting of goals*
4. The *opportunity*, in my management position, for *participation in the determination of methods and procedures*

V. Self-Actualization needs

1. The *opportunity for personal growth and development* in my management position
2. The *feeling of self-fulfillment* a person gets from being in my management position (that is, the feeling of being able to use one's own unique capabilities, realizing one's potentialities)
3. The *feeling of worthwhile accomplishment* in my management position

Procedure and Sample

A sample of 1,802 managers was obtained for this particular study by mailing the questionnaire to approximately 6,000 individuals holding management positions.³ (Some 100 additional individuals at the President level filled out the questionnaire, but their responses could not be used for this study because their organizational positions are neither specifically line nor specifically staff positions.) Approximately two thirds of the sample came from manufacturing companies and the remaining one third from nonmanufacturing organizations.

Respondents were cross-classified by line or staff type of position and by level of position within management. Line-staff classification was made on the basis of the respondent's self-classification of his position as either a line, a staff, or a combined line/staff position. The method of classification of respondents by vertical level of position has been previously described in detail (Porter, 1962). The four categories of level used in the present study were: Vice President, Upper-Middle, Lower-Middle, and Lower. This cross-classification by level was carried out for two reasons: to prevent the results for the line-staff variable from being confounded by the level variable (which has been shown previously to be strongly related to need satisfaction perceptions), and to provide four independent assessments of line-staff differences.

Table 1 presents the number of respondents in each of the three types of jobs—line, staff, and combined line/staff—within each of the four management levels. (This table can be referred to in determining the *Ns* for cells in the other tables of this article.) The table shows that there is very

³ The assistance of the American Management Association, and particularly Robert F. Steadman, in obtaining the sample of respondents is gratefully acknowledged.

TABLE 1

DISTRIBUTION OF *N* OF TOTAL SAMPLE BY LINE OR STAFF TYPE OF JOB AND FOUR LEVELS OF MANAGEMENT, AND CHARACTERISTICS OF SAMPLE BY TYPE OF JOB

Type of job	Management levels				Total <i>N</i> for job type	Median age	College degree (%)
	Vice President	Upper- Middle	Lower- Middle	Lower			
Line	186	187	138	41	552	42.9	74.9
Combined line/staff	227	177	94	15	513	44.1	71.8
Staff	198	295	199	45	737	42.2	75.7

little difference among the three major line-staff groups in regard to age or education. The combined line/staff group is slightly older, and the staff group has a slightly higher percentage of college graduates.

RESULTS

Deficiencies in Need Fulfillment

The method used to calculate perceived deficiencies in need fulfillment was to subtract the answer to Part a ("How much is there now?") from Part b ("How much should there be?") of each question for each respondent. With this method, the larger the difference between Parts b and a, the greater the perceived deficiency in fulfillment. The rationale of this measure, which is an indirect one derived from two direct answers of the respondents, has been explained previously (Porter, 1962).

The mean differences between Parts b and a for each item are presented in Table 2 for each subgroup of respondents. (Values for cells with an *N* of less than 20 are omitted in Table 2.) Two trends can be noted in Table 2: First, an examination of the 11 cells for each need item shows that the differences between the two ends of the horizontal dimension of organization structure—i.e., between line and staff jobs—are consistently and considerably smaller than the differences between the two ends of the vertical dimension of organization—i.e., between jobs at the Vice President level and jobs at the Lower management level. In other words, Table 2 shows that the variable of vertical level of position within management has a greater effect on perceived need fulfillment deficiencies than does the variable of line versus staff type of job. Secondly, Table 2

shows that for most of the items in the three highest-order needs, staff jobs produce greater deficiencies in fulfillment than do line jobs. However, certain items show no difference between the two types of jobs, and one item (IV-2) shows a definite reversal.

The trends for comparing the three different types of jobs are brought out more clearly in Table 3. In Table 3, changes in the size of mean deficiencies from line to combined line/staff to staff jobs are enumerated for each item and category. Whenever (in Table 2) a mean deficiency increased by more than .05 scale units from line to line/staff jobs or from line/staff to staff jobs for an item, an "increase" or + was counted; whenever a mean decreased from one column to the next (from the line column toward the staff column) by more than .05 scale units a "decrease" or — was counted. If changes were $\leq .05$ scale units, they were counted as "no change" and recorded as 0. Table 3 shows that two of the three Esteem items, two of the four Autonomy items, and all three of the Self-Actualization items produced definite trends for perceived need fulfillment deficiencies to increase from line to line/staff to staff jobs. By a simple sign test, the trend for line managers to be more satisfied than staff managers is significant in the Self-Actualization need area and approaches significance in the Esteem need area. In fact, for all 5 need areas and 9 of the 13 individual items, line managers were more often satisfied than staff managers where vertical level of position was held constant. For only 1 of the 13 items—"the opportunity for independent thought and action"—was a definite reverse trend present, with staff managers

TABLE 2
MEAN NEED FULFILLMENT DEFICIENCIES FOR EACH NEED CATEGORY ITEM
WITHIN EACH SUBGROUP OF RESPONDENTS

Need category	Item*	Management level	Type of job		
			Line	Combined L/S	Staff
Security	I-1	Vice President	0.38	0.49	0.45
		Upper-Middle	0.45	0.38	0.39
		Lower-Middle	0.37	0.40	0.38
		Lower	0.61	—	0.91
Social	II-1	Vice President	0.43	0.40	0.39
		Upper-Middle	0.41	0.36	0.52
		Lower-Middle	0.38	0.47	0.40
		Lower	0.63	—	0.40
	II-2	Vice President	0.04	0.30	0.18
		Upper-Middle	0.16	0.20	0.27
		Lower-Middle	0.28	0.15	0.26
		Lower	0.29	—	0.51
Esteem	III-1	Vice President	0.50	0.65	0.65
		Upper-Middle	0.75	0.90	0.92
		Lower-Middle	0.82	0.95	0.97
		Lower	1.12	—	1.58
	III-2	Vice President	0.46	0.33	0.47
		Upper-Middle	0.61	0.71	0.73
		Lower-Middle	0.69	0.82	1.02
		Lower	1.15	—	1.62
	III-3	Vice President	0.38	0.33	0.31
		Upper-Middle	0.51	0.38	0.37
		Lower-Middle	0.41	0.33	0.31
		Lower	0.46	—	0.71
Autonomy	IV-1	Vice President	0.64	0.67	0.66
		Upper-Middle	0.76	0.90	0.98
		Lower-Middle	0.88	0.94	1.07
		Lower	1.51	—	1.47
	IV-2	Vice President	0.64	0.44	0.41
		Upper-Middle	0.79	0.67	0.70
		Lower-Middle	0.93	0.84	0.80
		Lower	1.27	—	1.20
	IV-3	Vice President	0.62	0.72	0.66
		Upper-Middle	1.01	1.14	1.25
		Lower-Middle	1.17	1.32	1.33
		Lower	1.20	—	1.87
	IV-4	Vice President	0.38	0.40	0.41
		Upper-Middle	0.61	0.79	0.75
		Lower-Middle	0.74	0.76	0.68
		Lower	1.12	—	1.09
Self-Actualization	V-1	Vice President	0.85	0.92	0.98
		Upper-Middle	0.96	1.19	1.07
		Lower-Middle	0.86	1.39	1.16
		Lower	1.41	—	1.47

Table 2—Continued

Need category	Item ^a	Management level	Type of job		
			Line	Combined L/S	Staff
Self-Actualization	V-2	Vice President	0.80	0.84	0.91
		Upper-Middle	1.02	1.14	1.14
		Lower-Middle	1.15	1.10	1.21
		Lower	1.51	—	1.24
	V-3	Vice President	0.97	0.93	0.91
		Upper-Middle	1.03	1.16	1.30
		Lower-Middle	1.12	1.20	1.31
		Lower	1.29	—	1.62

^a For complete wording of items, refer to text.

being more satisfied than line managers. The conclusion, then, from Table 3, is that line managers tend to be more satisfied for almost all types of specific needs, with the single exception of chances to exercise independent thought and action. This general trend for line managers to feel that they are more satisfied in psychological need fulfillment

TABLE 3

NUMBER OF CHANGES IN SIZE OF MEAN DEFICIENCIES FROM LINE TO COMBINED LINE-STAFF TO STAFF TYPE OF JOBS WITHIN FOUR MANAGEMENT LEVELS

Need category	Item ^a	Management levels												Total sample		
		Vice President			Upper-Middle			Lower-Middle			Lower					
		+	0	—	+	0	—	+	0	—	+	0	—	+	0	—
Security	I-1	1	1	0	0	1	1	0	2	0	1	0	0	2	4	1
Social	II-1	0	2	0	1	1	0	1	0	1	0	0	1	2	3	2
	II-2	1	0	1	1	1	0	1	0	1	1	0	0	4	1	2
Category total		1	2	1	2	2	0	2	0	2	1	0	1	6	4	4
Esteem	III-1	1	1	0	1	1	0	1	1	0	1	0	0	4	3	0
	III-2	1	0	1	1	1	0	2	0	0	1	0	0	5	1	1
	III-3	0	2	0	0	1	1	0	1	1	1	0	0	1	4	2
Category total		2	3	1	2	3	1	3	2	1	3	0	0	10	8	3 ^b
Autonomy	IV-1	0	2	0	2	0	0	2	0	0	0	1	0	4	3	0
	IV-2	0	1	1	0	1	1	0	1	1	0	0	1	0	3	4
	IV-3	1	0	1	2	0	0	1	1	0	1	0	0	5	1	1
	IV-4	0	2	0	1	1	0	0	1	1	0	1	0	1	5	1
Category total		1	5	2	5	2	1	3	3	2	1	2	1	10	12	6
Self-Actualization	V-1	2	0	0	1	0	1	1	0	1	1	0	0	5	0	2
	V-2	1	1	0	1	1	0	1	1	0	0	0	1	3	3	1
	V-3	0	2	0	2	0	0	2	0	0	1	0	0	5	2	0 ^b
Category total		3	3	0	4	1	1	4	1	1	2	0	1	13	5	3*
Total all items		8	14	4	13	9	4	12	8	6	8	2	3	41	33	17**

^a For complete wording of items, refer to text.
^b Approaches significance ($p = .10$).
* $p \leq .05$.
** $p \leq .01$.

TABLE 4
MEAN IMPORTANCE FOR EACH NEED CATEGORY ITEM WITHIN EACH SUBGROUP OF RESPONDENTS

Need category	Item ^a	Management level	Type of job		
			Line	Combined L/S	Staff
Security	I-1	Vice President	5.44	5.37	5.52
		Upper-Middle	5.29	5.07	5.22
		Lower-Middle	5.25	5.36	5.29
		Lower	5.22	—	5.33
Social	II-1	Vice President	6.20	6.27	6.25
		Upper-Middle	6.09	6.01	6.18
		Lower-Middle	5.88	6.15	5.94
		Lower	6.00	—	5.91
	II-2	Vice President	4.63	4.64	4.75
		Upper-Middle	4.48	4.33	4.61
		Lower-Middle	4.71	4.50	4.79
		Lower	4.73	—	4.69
Esteem	III-1	Vice President	5.22	5.06	5.45
		Upper-Middle	5.18	5.24	5.33
		Lower-Middle	5.02	5.35	5.35
		Lower	5.17	—	5.29
	III-2	Vice President	5.53	5.36	5.47
		Upper-Middle	5.37	5.41	5.36
		Lower-Middle	5.36	5.55	5.46
		Lower	5.41	—	5.27
	III-3	Vice President	5.38	5.26	5.29
		Upper-Middle	5.13	5.27	5.14
		Lower-Middle	4.99	5.10	5.14
		Lower	5.00	—	5.04
Autonomy	IV-1	Vice President	6.16	5.78	5.63
		Upper-Middle	5.85	5.72	5.42
		Lower-Middle	5.72	5.62	5.29
		Lower	5.54	—	4.93
	IV-2	Vice President	6.50	6.37	6.33
		Upper-Middle	6.20	6.25	6.25
		Lower-Middle	6.06	6.18	6.07
		Lower	6.17	—	6.18
	IV-3	Vice President	6.38	6.35	6.23
		Upper-Middle	6.09	6.01	6.00
		Lower-Middle	5.95	5.82	5.78
		Lower	5.73	—	5.47
	IV-4	Vice President	5.91	5.84	5.71
		Upper-Middle	5.72	5.75	5.57
		Lower-Middle	5.67	5.73	5.36
		Lower	5.37	—	5.42

^a For complete wording of items, refer to text.

Table 4—Continued

Need category	Item ^a	Management level	Type of job		
			Line	Combined L/S	Staff
Self-Actualization	V-1	Vice President	6.41	6.16	6.33
		Upper-Middle	6.32	6.41	6.21
		Lower-Middle	6.28	6.30	6.30
		Lower	6.39	—	6.27
	V-2	Vice President	6.48	6.33	6.36
		Upper-Middle	6.23	6.28	6.33
		Lower-Middle	6.16	6.22	6.22
		Lower	6.37	—	6.24
	V-3	Vice President	6.59	6.46	6.50
		Upper-Middle	6.36	6.47	6.42
		Lower-Middle	6.20	6.30	6.30
		Lower	6.27	—	6.24

aspects of their jobs held up at all four management levels from Vice President down to Lower management.

It should also be noted from both Tables 2 and 3 that combined line/staff jobs are clearly intermediate between pure line jobs and pure staff jobs in terms of the sizes of perceived deficiencies. Line jobs more frequently produced smaller deficiencies than the combined line/staff jobs, and the combined line/staff jobs in turn more frequently produced smaller deficiencies than staff jobs. This intermediate position of the combined line/staff jobs lends additional support to the stability of the finding of differences in perceived deficiencies between the two ends of the horizontal dimension of organization structure. Apparently, one can speak of a meaningful continuum going from pure line jobs through jobs combining both line and staff features to pure staff jobs.

Importance of Needs

Table 4 is comparable to Table 2 and presents the mean importance attached to each item by each subgroup of respondents. The values in Table 4 were obtained by finding the mean answers to Part c ("How important is this to me?") of each item. This table shows that for all of the five need areas except that of Autonomy, there was little difference among the respondents in the three

types of jobs in terms of the importance they attached to the items in the different areas. That is, a given item was regarded as about equally important by the three groups of subjects. However, this was not true for the Autonomy items. This need area was regarded as more important by line managers.

Table 5, comparable in construction to Table 3, more clearly shows the absence of trends in the Security, Social, Esteem, and Self-Actualization areas, and the presence of a strong trend in the Autonomy area. The four Autonomy items, taken together, produced a highly significant trend for line personnel to attach more importance to this need than staff or combined line/staff managers. It should be noted that the Self-Actualization area, which produced a significant trend of greater perceived deficiencies in staff jobs (see Table 3), was regarded as about equally important by line and staff managers. Likewise, the Esteem area, which produced a trend approaching significance for the deficiency data, produced essentially no trend for the importance data. Thus, deficiency trends and importance trends tended to be relatively uncorrelated.

DISCUSSION

In recent years, some experts in the field of managerial and organizational behavior have contended that distinctions between line

TABLE 5
NUMBER OF CHANGES IN SIZE OF MEAN IMPORTANCE FROM LINE TO COMBINED LINE-STAFF
TO STAFF TYPE OF JOBS WITHIN FOUR MANAGEMENT LEVELS

Need categories	Item ^a	Management levels												Total sample		
		Vice President			Upper-Middle			Lower-Middle			Lower					
		+	0	-	+	0	-	+	0	-	+	0	-	+	0	-
Security	I-1	1	0	1	1	0	1	1	0	1	1	0	0	4	0	3
Social	II-1	1	1	0	1	0	1	1	0	1	0	0	1	3	1	3
	II-2	1	1	0	1	0	1	1	0	1	0	1	0	3	2	2
Category total		2	2	0	2	0	2	2	0	2	0	1	1	6	3	5
Esteem	III-1	1	0	1	2	0	0	1	1	0	1	0	0	5	1	1
	III-2	1	0	1	0	2	0	1	0	1	0	0	1	2	2	3
	III-3	0	1	1	1	0	1	1	1	0	0	1	0	2	3	2
Category total		2	1	3	3	2	1	3	2	1	1	1	1	9	6	6
Autonomy	IV-1	0	0	2	0	0	2	0	0	2	0	0	1	0	0	7**
	IV-2	0	1	1	0	2	0	1	0	1	0	1	0	1	4	2
	IV-3	0	1	1	0	1	1	0	1	1	0	0	1	0	3	4
	IV-4	0	0	2	0	1	1	1	0	1	0	1	0	1	2	4
Category total		0	2	6	0	4	4	2	1	5	0	2	2	2	9	17**
Self-Actualization	V-1	1	0	1	1	0	1	0	2	0	0	0	1	2	2	3
	V-2	0	1	1	0	2	0	1	1	0	0	0	1	1	4	2
	V-3	0	1	1	1	1	0	1	1	0	0	1	0	2	4	1
Category total		1	2	3	2	3	1	2	4	0	0	1	2	5	10	6
Total all items		6	7	13	8	9	9	10	7	9	2	5	6	26	28	37

^a For complete wording of items, refer to text.
** $p \leq .01$.

and staff positions are, or should be, diminishing. A related observation concerns the supposed increase in power of the staff relative to that of the line. In this latter connection, for example, McGregor (1960) has recently stated that, "indirectly, perhaps, but definitely and increasingly, the industrial organization is being run by the staff" (p. 156). Although such conclusions about line-staff relationships may be true, the findings of the present study show that a distinction between the two types of jobs, at least in terms of psychological need satisfactions, is still meaningful. It is apparent from the results presented in the preceding section that there are definite trends in line versus staff differences in perceived deficiencies in need fulfillment, and even some differences in perceived importance of needs. Considering all the areas of needs studied, line managers

feel they are more satisfied on their jobs than are staff managers. This is especially true in the Esteem and Self-Actualization need categories, two areas of needs to which both line and staff personnel assign about equal importance. In the Autonomy need area the two groups of managers express about the same degree of satisfaction, but line managers attach significantly more importance to these needs. Only a similar study carried out on a similar sample of line and staff managers at some point in the future will be able to prove conclusively whether these differences in need satisfaction and importance are decreasing, as some would expect, remaining the same, or increasing. Whatever the nature of the changes over time in the differences in perceived need satisfaction between managers in line versus staff positions, for the present it appears

to be necessary to consider this horizontal aspect of organization structure as one of the factors influencing job attitudes in management. However, it is also apparent from this and the previous study on levels of management that the horizontal dimension of organization structure has a weaker relationship to job attitudes than does the vertical dimension.

REFERENCES

- BROWN, A. Some reflections on organization: Truths, half-truths, and delusions. *Personnel*, 1954, 31(1), 31-42.
- FISCH, G. G. Line-staff is obsolete. *Harv. Bus. Rev.*, 1961, 39(5), 67-79.
- MCGREGOR, D. *The human side of enterprise*. New York: McGraw-Hill, 1960.
- MASLOW, A. H. *Motivation and personality*. New York: Harper, 1954.
- PORTER, L. W. A study of perceived need satisfactions in bottom and middle management jobs. *J. appl. Psychol.*, 1961, 45, 1-10.
- PORTER, L. W. Job attitudes in management: I. Perceived deficiencies in need fulfillment as a function of job level. *J. appl. Psychol.*, 1962, 46, 375-384.
- PORTER, L. W. Job attitudes in management: II. Perceived importance of needs as a function of job level. *J. appl. Psychol.*, 1963, 47, 141-148.

(Received August 6, 1962)

A VERIFICATION SCALE FOR THE MINNESOTA VOCATIONAL INTEREST INVENTORY¹

DAVID P. CAMPBELL AND RACHEL W. TROCKMAN

University of Minnesota

A verification scale, designed to detect individuals answering carelessly or incorrectly, was developed for the Minnesota Vocational Interest Inventory. The scale is composed of items answered very infrequently by Clark's group of Tradesmen-in-General. Data on a validation and cross-validation group are presented. The scale was shown to correctly identify 97% of arbitrarily responding individuals while misclassifying only 9% of individuals answering in a normal manner. To demonstrate other attributes of the scale, data are presented for a test-retest group ($r_{xx} = .81$), a hospitalized psychotic group, and a group of answer sheets completed using random numbers.

An index of whether or not an individual has responded to a psychological test or inventory in a careful, conscientious manner is frequently useful in the applications of such instruments, both in applied and research uses. Attempts to develop such indexes have been made for the Minnesota Multiphasic Personality Inventory (Hathaway & McKinley, 1945), the Strong Vocational Interest Blank (Filbeck & Callis, 1961), and the Kuder Personal Preference Record (Kuder, 1956). Usually these methods utilize items selected infrequently by most people, the rationale being that anyone who selects a substantial number of these low-frequency items is either being careless, does not understand the directions, or is deliberately trying to invalidate the results.

This article reports an attempt to develop a similar index for the Minnesota Vocational Interest Inventory (MVII).

The data used were those collected in the original validation of this inventory (Clark, 1961). Responses of the original group of Tradesmen-in-General (TIG) were scanned to find items chosen infrequently,² and five scales were developed using items chosen, respectively, by 5, 6, 7, 8, 9% or less by this group. (The scales were cumulative in the sense that the "5% or less" items all

appeared in the "6% or less" scale, etc.) The number of items in each scale is shown in Table 1.

When the scales were completed, the TIG validation (VAL) group was scored on them. Means and standard deviations for TIG-VAL are reported in Table 1. When items are selected in this manner, there is the considerable risk that the low frequency of response (and thus the selection of the item) is due solely to chance and would not be found in a similar group. To check this possibility, a second group of TIG was drawn from Clark's other criterion groups and used as a cross-validation (X-VAL) group. Means and standard deviations for TIG X-VAL are also reported in Table 1. The t tests between the means of the TIG-VAL and TIG X-VAL groups show these groups to be significantly different on all scales. There is some chance effect in the selection of these items, reflected by the significantly higher mean score of the X-VAL group.

Although the t tests indicate a real and consistent difference between groups, they do not indicate whether this cross-validation shrinkage is sufficient to seriously affect the practical utility of the scale. This latter decision is a practical, not a statistical one. In this case, the shrinkage is small. The differences between the VAL and X-VAL groups range from .7 to 2 points and are of dubious practical significance.

Another statistic which gives some insight about the practical magnitude of these dif-

¹ Support for this project came from the Numerical Analysis Center and the Graduate School at the University of Minnesota.

² Because Clark plans to shorten this inventory, only the first 96 triads were used.

TABLE 1
RESULTS OF SCORING SEVERAL GROUPS ON THE MVII's FIVE SCALES

Groups	N	Scales									
		5%		6%		7%		8%		9%	
		M	SD	M	SD	M	SD	M	SD	M	SD
TIG-VAL ^a	240	1.10	1.62	1.76	2.51	2.69	3.54	3.33	4.29	4.94	5.86
TIG X-VAL ^a	255	1.81	2.43	2.72	3.62	3.96	4.86	4.79	5.69	6.91	7.56
t VAL versus X-VAL		2.765**		2.47*		2.38*		2.31*		2.31*	
% overlap VAL versus X-VAL		98.6%		87.6%		88.0%		88.4%		88.3%	
Arbitrary respondents	27	8.33	3.02	11.67	3.49	16.04	3.90	19.15	3.89	24.41	4.33
Random number	30	8.36	2.61	11.66	2.69	16.00	2.85	18.40	1.98	24.40	1.93
Tradeschool test ^a	98	0.98	1.27	1.36	1.78	2.09	2.80	2.72	3.31	3.94	4.33
Tradeschool retest ^a	98	1.03	1.38	1.36	1.89	1.94	2.56	2.45	3.02	3.51	4.07
Test-retest reliability		.58		.72		.75		.81		.83	
Hospitalized psychotics ^a	16	3.94	3.89	6.94	5.69	9.56	7.72	11.38	9.06	15.88	11.59
Number of items on scale		26		37		50		58		76	

^a MVII answer sheets for these groups were made available by K. E. Clark.

* $p \leq .05$.

** $p \leq .01$.

ferences is the measure of percent overlap. This gives an indication of how many scores in one distribution can be matched by scores in another distribution. This statistic reveals considerable similarity between the VAL and X-VAL groups. The percent overlap, reported in Table 1, ranges from 87.6% to 98.6% and averages 90%. Using this approach, the keys hold up well under cross-validation. However, to be cautious, only the X-VAL group was used to determine how well these keys identify careless respondents.

To establish the validity of these scales, a group of people ($N = 27$) was given the answer sheet for the MVII (but not the booklet) and asked to respond blindly to the items. Means and standard deviations for this arbitrarily responding group are listed in Table 1. As can be seen, these means differ markedly from both the VAL and X-VAL TIG groups, in every instance by at least $2\frac{1}{2}$ standard deviations. The scales do identify people answering randomly.

Graphs were prepared for each key show-

ing the cumulative frequency curves of the TIG X-VAL group and the arbitrarily responding group to determine usable cutting scores for identifying careless respondents.

Although there were no extreme differences between these keys on this criterion of effectiveness, the 8% key achieved a slightly better separation between the groups than the other keys, and as it has respectable reliability (reported below) and has roughly the number of items that Clark recommends for the MVII scales (Clark, 1961, p. 37), it is the one that the authors recommend for use.

The graph for this key is reproduced in Figure 1. The dotted lines show that a cutting score of 13 will correctly identify 97% of the arbitrary respondents, while incorrectly classifying 9% of careful respondents. (Note that this 9% is a conservatively high figure; we have no assurance that everyone in Clark's criterion groups answered carefully.)

For a further comparison, 30 answer

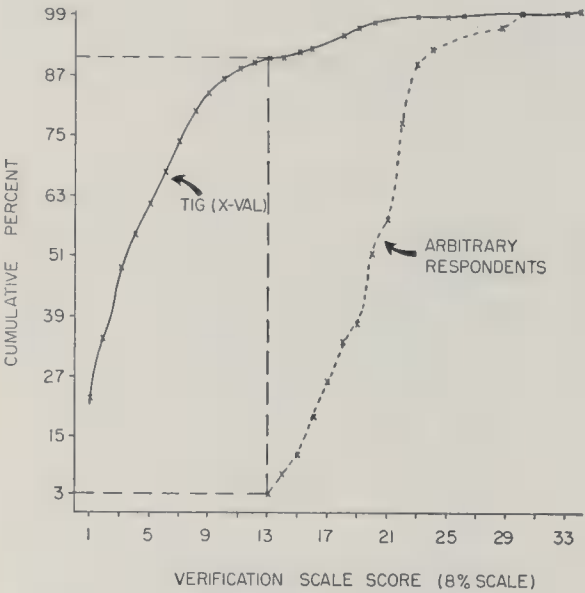


FIG. 1. Cumulative frequency distributions for TIG X-VAL and arbitrary respondents.

sheets were completed using a random number table. These were scored and the data are reported in Table 1. The random number means are extremely close to the means of the arbitrarily responding group, differing only in the second decimal place on four of the five keys. (In each instance, the standard deviation of the arbitrarily responding group is larger than the random number standard deviation, leading to the somewhat satisfying conclusion that people are more variable than random numbers.)

Reliability data were obtained by scoring a test-retest group. This was a group of students in a Minneapolis trade school who completed the MVII twice within 30 days. Frequently when gathering this type of test-retest data, investigators are concerned that individuals faced with the chore of responding again to a inventory that they have recently completed might be more careless the second time. Comparison of means on the Verification scale for this test and retest group should give some information on this point, and they are reported in Table 1. These data show no indication of careless responding on the second administration, a result which should mitigate but not eliminate fears of other investigators collecting analogous data in other situations.

Test-retest correlations are reported in

Table 1. The reliabilities increase as the scales lengthen and, for the two longest scales, fall with the range of reliabilities for the occupational scales of the MVII reported by Clark (1961, p. 41).

There is some possibility that a person might score high on these scales, even though responding carefully and honestly, because he happens to feel differently about these low frequency items. His likes and dislikes may be considerably different from those of the tradesmen used to develop the scale. Testmakers and users should be slow to label this kind of responding as undesirable in any sense. When a high score appears, this should be added to other data about the individual and the counselor or other user should proceed accordingly. Usually it will suggest that the inventory should be completed again, but not always.

To determine what the possible range of scores on these keys might be among groups very different from Clark's tradesmen groups, a group of hospitalized psychotics was scored. Although it was a small group ($N = 16$) and interpretations must be cautious, the data in Table 1 indicate this group falls somewhere between the tradesmen groups and the arbitrarily responding group. A slight majority fall below the recommended cutting score and thus would be considered to be answering the inventory carefully. However, the cutting score is less than $\frac{1}{4}$ standard deviation above the mean of this group so many of them would be found above this score. It is possible for individuals to score fairly high on this scale because they look at the world differently, not only because they are being careless. (Some readers may wish to interpret this data to indicate that hospitalized psychotics are simply more careless, an interpretation that still supports the use of this scale.)

It seems reasonable to conclude that a counselor seeing an individual with a high score on this scale should check to determine the reason before proceeding further. In other applications of the MVII such as research projects, etc., profiles with high scores on the Verification scale should be treated with suspicion. Unless it can be shown that it is

reasonable to expect the individual to differ widely in his interests from skilled tradesmen, the inventory should be readministered with special emphasis on completing it carefully and correctly.

REFERENCES

- CLARK, K. E. *Vocational interests of non-professional men*. Minneapolis: Univer. Minnesota Press, 1961.
- FILBECK, R. W., & CALLIS, R. A. Verification scale for the Strong Vocational Interest Blank, Men's Form. *J. appl. Psychol.*, 1961, **45**, 318-324.
- HATHAWAY, S. R., & MCKINLEY, J. C. *Manual for the MMPI*. New York: Psychological Corporation, 1945.
- KUDER, G. F. *Manual for the Kuder Preference Record, Form C*. Chicago: Science Research Associates, 1956.

(Received August 10, 1962)

TIME, AWARENESS, AND ORDER OF PRESENTATION IN OPINION CHANGE

DUANE P. SCHULTZ¹

American University

Primacy-recency was investigated as a function of time interval interpolated between the presentation of the opposing arguments and of the level of awareness of the manipulatory intent of the experimenter. Utilizing the pretest—experimental treatment—posttest design, 210 Ss were made differentially aware of the manipulatory intent of the experimenter. Half of the Ss were read the opposing arguments in immediate succession while the other half had a 2-week time interval interpolated between the argument presentations. Results indicated that those Ss who were unaware of manipulatory intent yielded a significant recency effect. Recency was minimized with those Ss who were made differentially aware of manipulatory intent.

PROBLEM

In the area of communication and persuasion, one question which has resisted resolution is whether the communication presented first is more effective in changing opinion in the direction of its argument or the communication presented last is more effective in changing opinion in the direction of its argument. This problem has been labeled the primacy-recency controversy in which primacy refers to the former effect and recency to the latter.

The studies undertaken in this area (Cromwell, 1950; Hovland & Mandell, 1952; Knowler, 1936; Lund, 1925) have yielded no agreement on the efficacy of either primacy or recency to consistently induce opinion change. This situation has led to the general conclusion that it may not be meaningful to postulate a universal law of either primacy or recency (Hovland, Mandell, Campbell, Brock, Luchins, Cohen, McGuire, Janis, Feierabend, & Anderson, 1957). Accordingly, recent research has been directed toward analyzing the various factors responsible for the greater effectiveness of the first or second communication.

Two variables which have been indicated for further study (Hovland, Janis, & Kelley, 1953; Hovland et al., 1957) are: (a) the time interval between the two opposing blocks of information, and (b) the level of

awareness on the part of the subjects of the intent of the experimenter to deliberately manipulate their opinions; i.e., do the subjects feel that the communicator is purposely trying to change their opinions? This study is an attempt to experimentally investigate the effect of these two variables on primacy-recency conditions of opinion change.

In the primacy-recency studies mentioned above, only a very short time interval elapsed between the two conflicting communications. Several studies (Luchins, 1957; Miller & Campbell, 1959; Underwood, 1948) have indicated that the greater the time interval between the two learning situations, i.e., the presentation of the opposing communications, the greater the operation of recency.

As a result of our normal developmental experiences as children in which we are frequently misled, disappointed and exploited by our parents and others for their own manipulative ends, we are apt to become highly sensitive to cues of manipulatory intent. Hovland and associates (1953) hypothesized that as a result of learning experiences of this kind, people would acquire strong motives to avoid being influenced when they expect that a communicator is consciously attempting to do so. It would seem then that expectations of manipulative intent are likely to give rise to strong tendencies to resist accepting the contents of a communication.

Thus, it is predicted that subjects who

¹ Now at Mary Washington College, University of Virginia.

are aware of the manipulatory intent of the investigator will yield no significant change from their pretest scores. If any opinion change is produced it will be in the direction opposite to the position advocated in the last communication; i.e., a primacy effect. Lack of awareness of manipulatory intent, then, should be directly related to a recency effect. In line with the evidence cited above on the effect of a time delay between the opposing communications, it is felt that the interpolation of the 2-week time interval may facilitate a recency effect.

METHOD

A five-item Likert-type questionnaire was used as the pretest and posttest in order to evaluate opinion toward the topic in question. Each item contained five alternative choice responses ranging from full agreement with the item statement through "neutrality" to full disagreement with the statement. A high score value represented the pro position while a low score value represented the con position. The topic involved the problem of how much effort, in time and money, should be devoted to cancer research in the next 5 years. A pro and a con argument developed by Hovland and associates (1953) were used as the persuasive materials in the study.

The positive argument stressed the likelihood of a cure for cancer within the next 5 years if a determined effort is put into effect. The con position stated the hopelessness of finding a cure for cancer in the near future and urged that effort be applied to finding cures for other diseases for which the outlook is more optimistic.

The two levels of awareness instructions were read to certain of the groups in order to inform them that the purpose of the study was to deliberately manipulate or change their attitudes. The "strongly aware" and "moderately aware" instructions follow.

Strongly Aware Instructions:

The purpose of this experiment is to attempt to change your attitudes on the topic: A cure for cancer—how soon? Time and time again the communication procedures utilized here have been shown to be capable of changing the attitudes of college student subjects. The method has proven so successful that there seems to be little doubt that your attitudes will be influenced.

Moderately Aware Instructions:

The purpose of this study is to determine the attitudes of college students on the topic: A cure for cancer—how soon? It is felt that there is some chance that the attitudes of college students might be capable of being influenced somewhat by the procedures utilized here. The main purpose of the study, however, is to determine your attitude on the topic.

A total of 210 undergraduates at the American University served as subjects in this experiment. The subjects were randomly assigned to 1 of 12 treatment groups.

All subjects were administered the pretest questionnaire to determine their pretreatment attitudes on this topic. Group A1 was made strongly aware of the manipulative intent of the experiment by the strongly aware instructions delivered prior to the presentations of the conflicting arguments. Group A2 was made moderately aware of the manipulative intent of the study by the moderately aware instructions again delivered prior to the argument presentations. Group A3 was given no instructions nor in any way informed of any manipulative intent on the part of the communicator.

Each of these A groups was then divided into two subgroups: B1 and B2. A 2-week time interval was interpolated between the presentation of the conflicting arguments for Subgroup B1, while Subgroup B2 received both communications at the same time, one immediately following the other.

Each of the B1 and B2 subgroups was then further divided into two subgroups: C1 and C2. Subgroup C1 received a pro-con order of presentation of the arguments while Subgroup C2 received a con-pro order. Finally, the posttest questionnaire

TABLE 1
EXPERIMENTAL DESIGN OF PRIMACY-RECENCY STUDY

A1 Strongly aware				A2 Moderately aware				A3 Not aware			
B1 2-week interval		B2 No interval		B1 2-week interval		B2 No interval		B1 2-week interval		B2 No interval	
C1	C2	C1	C2	C1	C2	C1	C2	C1	C2	C1	C2
Pro	Con	Pro	Con	Pro	Con	Pro	Con	Pro	Con	Pro	Con
Con	Pro	Con	Pro	Con	Pro	Con	Pro	Con	Pro	Con	Pro

Note.—Pretest and posttest questionnaires were administered prior to and after treatment.

TABLE 2
ANALYSIS OF COVARIANCE FOR ALL
SOURCES OF VARIATION

Source	df	MS	F
Order (C)	1	145.984	46.227*
Time (B)	1	0.460	0.146
Awareness (A)	2	2.886	1.828
C × B	1	9.471	2.999
C × A	2	46.505	14.726*
B × A	2	1.904	0.602
C × B × A	2	0.000	
Error	168	3.158	

* $p \leq .05$.

was administered immediately after the presentation of the second argument. The design is summarized in Table 1.

RESULTS

The data were first treated by a three-way analysis of covariance, the summary of which is contained in Table 2.

Inspection of Table 2 reveals that two of the sources of variation are significant at the .05 level: order effects (C) and the Order × Awareness interaction effect (C × A).

The data were then analyzed by using the subtractive procedure described by Hovland and Mandell (1952). Pretest means were subtracted from their respective posttest means in each C1 and C2 subgroup. Then each C1 difference score was subtracted from its C2 score. This yielded two "difference between differences" scores, one for each B1 and B2 subgroup. These final difference scores were then tested against the null hypothesis by the use of a t test. If the numerator of

the t ratio is positive then a primacy effect is demonstrated (if statistically significant). If the numerator is negative, a recency effect is shown (if statistically significant).

A summary of the mean pretest and posttest scores for all groups and the results of the t tests are found in Table 3.

Inspection of Table 3 reveals that significant order effects were found in Group A3 and in the 2-week interval subgroups of Group A2. With the group not made aware of manipulatory intent, significant recency effects were found in both B1 and B2; i.e., both the group with the 2-week time interval between opposing communications and the group with no time interval.

With the group made moderately aware of manipulatory intent a significant primacy effect was found only in that subgroup with the 2-week time interval. The group made strongly aware of manipulatory intent yielded no significant order effects under either of the time interval conditions.

DISCUSSION

The results of the analysis of covariance clearly indicate that there were significant differences among the experimental groups with respect to order effects. The significant C × A interaction effect indicates that differential effects on attitude change were produced by the experimental manipulation of the variables of level of awareness of manipulatory intent and order of presentation. Thus the variable of time delay was not revealed to be a significant factor affecting

TABLE 3
SUMMARY OF PRETEST AND POSTTEST MEAN SCORES, DIFFERENCE SCORES, AND t VALUES

	A1				A2				A3			
	B1		B2		B1		B2		B1		B2	
	C1	C2	C1	C2	C1	C2	C1	C2	C1	C2	C1	C2
Pretest	17.33	18.93	17.13	17.40	19.20	19.26	18.26	17.33	18.60	17.40	18.66	17.13
Posttest	16.40	18.06	15.86	16.33	16.73	18.66	16.80	16.80	16.86	17.73	14.80	19.40
Difference score												
t												

Note.—df = 58.
* $p < .05$.

conditions of opinion change, at least in interaction with level of awareness.

Examination of the t values in Table 3 indicates support of the prediction on the effects of level of awareness; i.e., lack of awareness of manipulatory intent yielded a significant recency effect.

The results were not significant under the strongly aware condition. However, the moderately aware group yielded a significant primacy effect with the 2-week time interval. There was no significant directional effect in the no-time-interval subgroup.

These findings are consistent with the results of a study by Lana (1961) who found that increased familiarity with the issue increased primacy effects. Lana created experimentally differential levels of familiarity much in the same way that differential levels of awareness were created in this study. Lana found that those subjects who received no information other than that contained in the pro and con arguments exhibited a significant recency effect. In the present study, those subjects who received no information other than that contained in the opposing arguments, i.e., were not made aware of manipulatory intent, also yielded significant recency effects. Recency was minimized with those subjects who were made differentially aware of the manipulatory intent of the study.

The lack of agreement of the time delay variable with the previous relevant studies (Luchins, 1957; Miller & Campbell, 1959; Underwood, 1948) remains for further investigation.

That subjects may resist accepting the

content of a countercommunication when they are aware of a deliberate attempt to change or manipulate their opinions is certainly of practical and theoretical value in the study of communication and persuasion.

REFERENCES

- CROMWELL, H. The relative effect on audience attitude of the first versus the second argumentative speech of a series. *Speech Monogr.*, 1950, 17, 105-122.
- HOVLAND, C. I., JANIS, I. L., & KELLY, H. H. *Communication and persuasion*. New Haven: Yale Univer. Press, 1953.
- HOVLAND, C. I., & MANDELL, W. An experimental comparison of conclusion drawing by the communicator and by the audience. *J. abnorm. soc. Psychol.*, 1952, 47, 581-588.
- HOVLAND, C. I., MANDELL, W., CAMPBELL, ENID H., BROCK, T., LUCHINS, A. S., COHEN, A. R., MCGUIRE, W. J., JANIS, I. L., FEIERABEND, ROSALIND L., & ANDERSON, N. H. *The order of presentation in persuasion*. New Haven: Yale Univer. Press, 1957.
- KNOWER, F. H. Experimental studies of changes in attitude: II. A study of the effect of printed argument on changes in attitude. *J. abnorm. soc. Psychol.*, 1936, 17, 315-347.
- LANA, R. E. Familiarity and the order of presentation of persuasive communications. *J. abnorm. soc. Psychol.*, 1961, 62, 573-577.
- LUCHINS, A. S. Primacy-recency in impression formation. In C. I. Hovland et al. (Eds.), *The order of presentation in persuasion*. New Haven: Yale Univer. Press, 1957. Pp. 33-61.
- LUND, F. H. The law of primacy in persuasion. *J. abnorm. soc. Psychol.*, 1925, 20, 183-191.
- MILLER, N., & CAMPBELL, D. T. Recency and primacy in persuasion as a function of the timing of speeches and measurements. *J. abnorm. soc. Psychol.*, 1959, 59, 1-9.
- UNDERWOOD, B. J. Retroactive and proactive inhibition after five and forty-eight hours. *J. exp. Psychol.*, 1948, 38, 29-38.

(Received August 13, 1962)

RELATIONSHIPS BETWEEN CARBON CHAIN LENGTH AND AVOIDANCE RESPONSES IN RATS¹

FRANCIS J. BLAISDELL

International Electric Corporation, Paramus, New Jersey

An experiment was performed to ascertain the results of subjecting the conditional twine cutting behavior of 32 white rats to the even members of a homologous series of equimolar solutions of amine hydrochlorides. The Shurrager multipartitioned training cage (5-barrier) was used to train the animals to constant speed on untreated barriers. The middle barriers were later used randomly in a balanced design as test barriers. Results show the order of repellency of the series (from most to least): $12 > 10 > 16 > 18 > 14 > 8 > 6 > 4$. Conclusion: increasing carbon chain length produces a maximum repellency at C_{12} , which is significantly different ($.02 \leq p \leq .05$) from C_4 , C_{10} , and C_{16} .

The behavior of the laboratory rat toward acceptable incentives has had considerable emphasis in psychological research. Many instrumental learning studies have required animals to undergo states of deprivation to produce motivated behavior toward food, water, or other objects as incentives. Other instrumental learning experiments have employed electric shock while an animal was under strong motivation in order to elicit aversion reactions. In general, the range of possible negative or repelling objects studies, other than electric shock, is rather narrow.

In a recent investigation of baler twine cutting by rats, conducted by the Armour Research Foundation (1954) of the Illinois Institute of Technology, Shurrager suggested some new methods of approach to the problem of repellents. One special apparatus, the Shurrager multipartition cage, was developed. It enabled the experimenter to observe the

rat's cutting behavior and at the same time to make accurate time recordings of such behavior. The method thus provided measurable aversion reactions to repellents which might be placed on the twine after the animal's cutting response had been well established.

The other device developed for the Armour Research Foundation study was the Shurrager matrix apparatus, designed specifically to detect the order in which an animal might cut treated twines having various repellent chemicals on them at any one barrier or choice point. The recording or order of choice and time was by means of a kymograph.

The Armour Research Foundation (1954) investigation had as its purpose the search for absolute repellents against twine cutting in rats, laboratory and field. That investigation, however, did not focus on any particular class of repellents for systematic emphasis. There were no attempts to study, for example, hereditary and sex differences which would be of psychological and genetic interest. Shurrager,² however, pointed out that:

individual differences in rat's behavior toward repellents of various kinds at barriers strongly suggested hereditary differences in taste, odor, etc., in rats. This would make it difficult to find a single chemical as an absolute repellent—more than likely it would require a complex of chemicals for effective universal repellency.

In the study reported here, only one of the general methodological approaches devised by

¹ Submitted in partial fulfillment for the PhD degree in the Department of Psychology and Education, Illinois Institute of Technology.

The suggestion for this paper and the dissertation proposal is that of Phil S. Shurrager to whom I express my sincere thanks. The chemical specifications were devised by Leon Gershbein, Biochemical Laboratory, Illinois Institute of Technology.

Additional material has been deposited with the American Documentation Institute. Order Document No. 7563 from ADI Auxiliary Publications Project, Photoduplication Service, Library of Congress, Washington 25, D. C. Remit in advance \$1.25 for microfilm or \$1.25 for photocopies, and make checks payable to: Chief, Photoduplication Service, Library of Congress.

² P. S. Shurrager, personal communication, 1962.

Shurrager, the Shurrager multipartitioned cage, and only one class of chemicals was selected. The problem in this study was, specifically, to discover the effects of increasing the carbon chain (straight) length, for the even members of the class of amine hydrochlorides, on a learned twine cutting task for two strains of white rats.

METHOD

Subjects

The subjects were 36 white laboratory rats weighing 90 grams prior to handling and training. The rats consisted of two strains: the Frank's strain of which nine were females and nine were males; and the Sprague-Dawley strain of which nine were females and nine were males. They were given an optimum for about 2 weeks in which food was continually in the cage. During this period animals were maintained on Purina chow. Before the animals were placed in their individual cages, all 36 animals were allowed to explore daily the experimental cage at random, nine at a time. Moist Purina chow, finely ground, was placed in the final or goal cage. None of the doors to the experimental cage was closed since each was held up by a spring allowing free access to any compartment. During this 1-week exploratory period, the animals were gentled by the experimenter about 5 minutes per day.

Next, each animal was placed in a separate cage and the cage identified by strain, animal number, and sex. The animals remained so identified throughout the training and testing periods, as well as control trials. Around the twelfth week of age, the animals were shifted from Purina chow to Rockland mouse diet, R-2 pellets, manufactured by A. E. Staley Manufacturing Company, Decatur, Illinois. This was the only diet available to the animals for the remaining training and testing periods except for the goal box where ground Purina chow was always available.

Apparatus

The dimensions of the Shurrager multiple partitioned cage are $6 \times 1 \times 1$ feet. The top and sides are of Plexiglas and the bottom is wire screening. The cage is divided into six compartments by sheet aluminum partitions. There is a small door in each partition, attached at one end to a spring and at the other end of the door to the twine, which is held under tension by a 1 pound weight. The twine is stretched free from the partition so that it is completely accessible to the animal. By cutting the twine, the rat releases the door, which retracts upward because of the pull of the spring. This action reveals an opening to the rat and the animal advances to the next closed door. The ends of the Shurrager multipartitioned cage have sliding doors which are operated manually to allow the animal

to enter. (It is also possible to place the animal directly in any one compartment to begin a session.) All the spring operated doors, separating the five partitions, are operated by the action of the animal on the twine. The experimenter first ties a knot on one end of the twine. The knot is slung over a forked prong on the door. The door is then pulled closed. The spring attached to one end of the door is thereby stretched, causing tension, and the twine is looped to a pound weight beneath the screened bottom of the cage. There the twine is again looped around an extension of a microswitch which is held closed by the twine, and which suspends a freely hanging pound weight thereby completing the circuit. The twine is prevented from slipping by using a ball-bearing clip which holds the twine tightly to the metallic base adjoining each half of the cage. When the twine is completely cut through, the weight falls and the animal may proceed to the next compartment.

The cutting time is recorded automatically in seconds. A microswitch wired to an electrically motivated counter is positioned to each partition in such a way that contact is broken when the animal cuts through the twine and the weight falls causing the next counter to begin immediately.

Removable glass partitions divide the compartments into two sections, so that the two animals can be tested at the same time, and the efficiency of experiments are thereby doubled. If the glass partition is removed, the experiment may be so devised that a rat may have a choice between two repellents, two concentrations of the same repellent, or treated and untreated twine. Along the top of the cage are fluorescent lights which serve both to obscure the observer from the rat and to provide illumination for photography.

Procedure

Training in a single untreated cotton string. The 90-day-old animal was placed in the starting cage and the timing switch was begun immediately. At first, a dab of wet Purina chow was placed on the cotton twine near the place at the bottom of the wire cage where the twine led to the microswitch. This bait was satisfactory for some animals but not for others. Different baits, such as hamburger, cheese, butter, and a moistened Purina pellet were used to motivate cutting behavior. As training advanced, the bait was placed behind the closed partitioned door and the animal allowed to cut the string before being fed. Once the twine was cut, the hungry animal generally proceeded to the next cage quickly.

When the animal appeared to be learning to cut each of the five strings in succession and not each string as a separate problem, the placing of bait behind each door and on each string was stopped and only the wet Purina ground chow was used in the goal cage. The animal was allowed to eat in the goal cage about 2 or 3 minutes throughout all training and testing trials.

All 36 animals were given 24 training trials on the single-strand cotton string. It was observed that only a few animals could cut the string cleanly at first. Many animals preferred to thread the twine to pieces until it broke under the tension of the springed door. As time advanced, the animals became stronger and their teeth became more suitable to cutting string. Their general stance was to grab the twine with the forepaws and to brace themselves with their hindfeet and to lean against the glass partition. Some crouched and braced themselves while cutting. As the learning progressed the animals braced themselves to a lesser extent and merely attacked the twine directly with their teeth.

The rat's cutting behavior appears to be a fast, vibrating motion with the jaws as the teeth cut the twine. The good and medium cutters used this type of response but some animals preferred a grinding action by the jaws. No detailed record was made by sex or strain of these chewing movements. There were observed different classes of positioning and attack on the twine. It was observed that animals would orient themselves: (a) with the head to the left or right of the twine and cut at the base of the wire cage; (b) with the head left or right of the twine but cut preknot; (c) with the head left or right but cut postknot, the loop of the knot; (d) one rat, later discarded, shredded the twine until it broke; and (e) one rat, later discarded, actually untied the knot. Almost 80% of the animals oriented themselves by a and b.

Training on untreated binder twine. There was no general departure in training for untreated binder twine from that used for the cotton twine except that no baiting was necessary. Since the binder twine was tougher and thicker in diameter (varying from 100 to 150 strands per twine, according to International Harvester estimates), the minimum time to cut increased for the early trials for training over the later trials for the cotton twine. Fifty trials were given on binder twine.

Training on untreated double-strand cotton string. Because of the failure of the treated binder twine to withstand the tension (after being treated with the amine hydrochlorides) that was required to hold the barrier closed, it was necessary to revise the original plan to use the sisalana (baler) twine and instead to use double-strand cotton string to hold the door closed. The cotton string easily withstood the tension. Three trials were given all rats on the double-strand cotton string prior to testing.

The animals were run daily throughout training and testing. The total period covered about 6 months; hence, the animals were actually over-trained. The testing and control trials took only 11 days.

Stimulus series. The even-numbered homologous series, which was applied to untreated twine is listed in Table 1. In all cases, the concentration of the solution was 2.0 mols. All twines held an equal number of molecules of amine hydrochlorides.

These chemicals were then randomized by a procedure given by Williams (1949) so that each of the eight chemicals preceded and followed any one and all of the others only once. This resulted in a randomized, balanced Latin square for presentation order of an even-numbered homologous hydrochloride series. Beeswax, paraffin wax (Paraseal), and Ivory soap were used for control tests.

Since the Shurrager multipartitioned cage contained five barriers, and since only Barriers 2, 3, and 4 were to be treated, these barriers were also randomized. Since, if n barriers are odd, it is impossible to balance such a table without repeating the entire tests in reverse according to Williams, an unbalanced Latin square was used. Barrier 2 was treated 21 times, Barrier 3 was treated 22 times, and Barrier 4 was treated 21 times, for eight subjects per sex strain for eight trials. This controlled for animals which might learn to expect that a certain barrier would always be the treated barrier.

Since there were only eight chemicals, only eight animals per sex strain were used during testing and control sessions. Excess animals were removed by chance procedure.

The beginning and ending times were automatically recorded for each barrier and the difference between cutting one barrier and the cutting of the next barrier constituted the total response time for each barrier.

Method of application and presentation. Since the amine hydrochloride series was prepared in equimolar solutions, the double-strand cotton twine was dipped in the proper member of the series for 10 minutes at 60 degrees Centigrade removed, and placed on a drying rack for 24 hours. Then the randomized presentation of the various carbon lengths on the cotton twine barriers was given to the assigned rat for the day. On any one day, only one chemical per rat was tested. The total number of trials per rat was eight. No animal was tested more than once or any one chain length; however, each chain length was administered to 32 different animals.

RESULTS

During training sessions the usual negative accelerated performance curves were obtained. The Frank strain animals stabilized mean running time at 8 seconds while the Sprague-Dawley strain stabilized at 17 seconds. This difference was significant at the 1.98 level. During test trials, however, there was no significant difference between strains for running time on the untreated barriers.

Test results showed that Carbon-12 (C_{12}) repelled the greatest percentage of animals and C_4 the least. Ivory soap, used as a control chemical, did better than C_4 . Detailed

results are indicated in Table 1. In general, results show that as carbon chain length of the amine chlorides increases for even members of the carbon series, the number of animals avoiding the repellent increases, becomes a maximum at a carbon chain length of 12, and decreases sharply thereafter. Less than half the animals were repelled by chain lengths of C₄, C₆, C₁₄, C₁₆, and C₁₈.

A chi square test of the significance of the frequencies between sex and between strains produced a firm acceptance ($p < .01$) of the null hypothesis that there were no strain or sex differences in the sensitivity to the chemical repellents.

TABLE 1

FREQUENCY AND PERCENTAGE OF RATS REPELLED ON TEST TRIALS ACCORDING TO CARBON CHAIN LENGTH OF AMINE CHLORIDES AND CONTROL SUBSTANCES

Carbon chain length	Number ^a of animals repelled	Percent
4	10	31.25
6	14	43.75
8	16	50.00
10	19	59.37
12	22	68.74
14	13	40.62
16	14	43.75
18	13	40.62
Beeswax	8	25.00
Petroleum wax	3	09.37
Ivory soap	12	37.50

^a Total number of animals was 32.

When the behavior of the animals toward the treated and untreated barriers during experimental trials is compared, data indicate that the treated barriers did not slow up the animals' crossing the untreated barriers. This was true whether or not the untreated barriers came before or after the treated barriers and regardless of chain length. The untreated barriers were crossed uniformly at about 20 seconds, before or after crossing a treated barrier and regardless of chain length. Table 2 shows the data in detail.

Results also indicated that the relative positions of the carbon chain lengths remain

TABLE 2

COMPARISON OF ANIMALS' RUNNING TIME FOR TREATED AND UNTREATED BARRIERS BY CARBON CHAIN LENGTH

Carbon chain length	Untreated before ^a	Barriers after ^a	Treated barrier ^a
4	19	21	130
6	19	20	163
8	19	19	164
10	18	20	210
12	21	20	230
14	23	19	151
16	20	17	170
18	18	16	160

^a Mean running time in seconds for 32 animals.

stable in repelling the animals. The data are summarized in Table 3. The third column, mean z score (time), was computed by using each animal as its own control. The individual z score was the difference between absolute value of running time for test barrier minus the mean running time for untreated barrier for a given length, then this value was divided by the standard deviation for untreated barriers for a given animal. The mean z score was obtained from the individual z scores. Hence, by three statistical treatments, the relative positions of the carbon chain lengths remained the same.

The question arose as to whether the ap-

TABLE 3

INDICATION OF STABILITY OF POSITION OF AMINE CHLORIDE CHAIN LENGTH IN REPELLING ANIMALS

Chain length	Total animals repelled ^a (%)	Mean repelling time for all animals ^a	Mean z score ^b
4	31.25	130	15.67
6	43.75	163	16.86
8	50.00	164	17.29
10	59.37	210	25.41
12	68.74	230	29.06
14	40.62	151	18.69
16	43.75	170	20.52
18	40.62	160	19.09

^a $N = 32$.

^b In seconds.

parent magnitude of differences between the various mean z scores was significant. Since the frequency distribution of repellency times was definitely not normal the sign test (Siegel, 1956) was used to compare all chain lengths with Carbon Chain Length 12. The specific question asked was whether or not it is probable (at some level of significance) that C_{12} 's average z score is higher than some other given carbon chain length. A one-tailed test showed that C_{12} is superior in efficiency, significantly ($.02 \leq p \leq .05$) for three cases ($C_{12}-C_4$, $C_{12}-C_{10}$, and $C_{12}-C_{16}$) out of the seven possible comparisons. The sign test also showed that the order of repellency from most to least was: $12 > 10 > 16 > 18 > 14 > 8 > 6 > 4$. There was no significant difference between sex or strain in the frequency of absolute repellency.

DISCUSSION

The null hypothesis was: if the even homologous series of equimolar solutions of amine hydrochlorides produces no effect on the conditioned twine cutting responses, then there will be observed no significant difference between the cutting time for untreated barriers as compared with the cutting time for the treated barriers. The evidence showed that the null hypothesis was not supported ($p = .01$). The data show that the animals were definitely slowed down by the application of the amine hydrochlorides. The evidence as shown by the z score analysis and sign test showed that the most efficient repellent was produced by Carbon Chain Length 12 of this amine hydrochloride.

One could consider the situation for the rats as a "stress situation." After they had

been able to cut nontreated barriers to a high degree of skill, physiological aversion to the twine cutting response developed possibly because of the tactile cues from, and olfactory cues emanating from, the treated barriers. Curiously enough, some high performing animals would cut the treated barrier as fast as possible and then were so disturbed that, even though they cut succeeding untreated barriers on that run, they were observed to delay 3 or 4 minutes in eating at the goal box. This situation was especially true for C_{12} . When these animals returned to their home cages, excessive drinking, generally reddened lips and noses were noted. Excessive salivation and wetness of neck fur were observed, probably caused by the attack of the chemical on the throat producing difficulty in swallowing. Animals, after cutting C_{10} or C_{12} , would spend considerable time in the goal box wiping their feet rapidly on the bottom of the cage before eating.

Hence, this stress situation produced "divergent" reaction groups. Some animals cut straight through with their original proficiency; other animals cut only nontreated barriers and waited for the experimenter to cut the barrier at 300 seconds, the criterion.

REFERENCES

- ARMOUR RESEARCH FOUNDATION. *Baler twine: Its protection and preservation from rodent attack*. (Project No. C 522) Chicago: Illinois Institute of Technology, Technology Center, 1954.
- SIEGEL, S. *Nonparametric statistics for the behavioral sciences*. New York: McGraw-Hill, 1956.
- WILLIAMS, E. J. Experimental designs balanced for the estimation of residual effects of treatments. *Aust. J. scient. Res.*, 1949, 2, 149-168.

(Received September 7, 1962)

Journal of Applied Psychology

MARYGROVE COLLEGE LIBRARY
DETROIT, MICHIGAN

KENNETH E. CLARK, Editor
University of Rochester

PLEASE DO NOT REMOVE

Table of Contents

Pursuit Tracking with Differentiating and Integrating Control Systems: E. C. Poulton.....	289
Factors Affecting Perceptual Integration of Illustrated Material: James M. McKendry, Monroe B. Snyder, and Stephen Gates.....	293
Some Personality and Behavioral Factors Related to Birth Order: Ewart E. Smith and Jacqueline D. Goodchilds.....	300
Individual Differences in Selection Decisions: Patricia M. Rowe.....	304
A Factor Analysis of Experimental Social Desirability and Response Set Scales: Allen L. Edwards.....	308
A Modified Model for Test Validation and Selection Research: Marvin D. Dunnette.....	317
Emotional Arousal and Task Performance: Bibb Latané and A. John Arrowood.....	324
Dependency Responses to Televised Instruction: Shepard A. Insel, Kurt Schlesinger, and Wilfred Desrosiers.....	328
Defining the Perceived Functions of Purchasing Personnel: J. C. Denton and Erich P. Prien	332
Leader Assumed Dissimilarity as a Measure of Prejudicial Cognitive Style: Robert C. Ziller	339
Active Responding in Programed Learning Materials: Thomas F. Hartman, Barbara A. Morrison, and Margaret E. Carlson.....	343
Predicting Credit Risk with a Numerical Scoring System: James H. Myers.....	348

American Psychological Association

Consulting Editors

GEORGE E. BRIGGS, *Ohio State University*
NORMAN FREDERIKSEN, *Educational Testing Service*
LEONARD D. GOODSTEIN, *University of Iowa*
EDWIN R. HENRY, *Peace Corps*
JOHN HOLLAND, *American College Testing Program*
CLIFFORD E. JURGENSEN, *Minneapolis Gas Company*

LAURENCE S. MCGAUGHRAN, *University of Houston*
QUINN MCNEMAR, *Stanford University*
HAROLD F. ROTHE, *Beloit Corporation*
THOMAS A. RYAN, *Cornell University*
ALEXANDER G. WESMAN, *Psychological Corporation*
CLARK L. WILSON, *Batten, Barton, Durstine, and Osborn, Incorporated*

This journal gives primary consideration to original investigations in any field of applied psychology except clinical psychology, although a descriptive or theoretical article may be accepted if it represents a special contribution in an applied field. Quantitative investigations of interest or value to psychologists working in the following broad fields will be considered: vocational and educational prognosis, diagnosis, and guidance at the secondary and college level; personnel research in business, industry, and government; engineering psychology; industrial working conditions; research on opinion and morale factors; job analysis and classification research; market and advertising research.

Manuscripts must be accompanied by an abstract of 100–120 words typed on a separate sheet of paper. The abstract should conform to the style of *Psychological Abstracts*. Detailed instructions for preparation of the abstracts appeared in the *American Psychologist* (1961, 16, 833), or they may be obtained from the Editor or from the APA Central Office.

Manuscripts should be addressed to the Editor:

Dr. Kenneth E. Clark
College of Arts and Science
302 Morey Hall
University of Rochester
Rochester, N. Y. 14627

All manuscripts must be submitted in duplicate. Original figures are prepared for publication; duplicate figures may be photographs or pencil-drawn copies.

Manuscripts must conform to the style requirements described in the *Publication Manual of the American Psychological Association*.

The following policies govern reprinting of materials copyrighted in APA journals: (a) to require approval to reprint for tables and figures, and for text only if more than 500 words [total from one article] in length; (b) to continue approval to reprint, contingent on the author's approval, for articles reprinted in whole or in major part; (c) to negotiate where possible the dedication of royalties from commercial publishers to the American Psychological Foundation.

Journal of Applied Psychology

Published bimonthly by the
American Psychological Association
Prince and Lemon Sts., Lancaster, Pa. 17604
and 1333 Sixteenth St. N. W.
Washington, D. C. 20036

\$10.00 per volume

\$2.00 per issue

HELEN ORR
Managing Editor

ELIZABETH S. REED
Advertising Manager

VIRGINIA RICHARDS
Editorial Assistant

Subscriptions, orders, and business communications should be addressed to the American Psychological Association, 1333 Sixteenth St. N. W., Washington, D. C. 20036. Address changes must reach the subscription office by the tenth of the month to take effect the following month. Undelivered copies resulting from address changes will not be replaced; subscribers should notify the post office that they will guarantee second-class forwarding postage. Other claims for undelivered copies must be made within four months of publication.

Second class postage paid at Lancaster, Pa. and at additional mailing places.

© 1963 by the American Psychological Association, Inc.

Journal of Applied Psychology

VOL. 47, No. 5

OCTOBER 1963

PURSUIT TRACKING WITH DIFFERENTIATING AND INTEGRATING CONTROL SYSTEMS¹

E. C. POULTON

Applied Psychology Research Unit, Cambridge, England

A differentiating amplifier in the control loop, an integrating amplifier, and a simple amplifier were compared experimentally on random inputs with upper cutoffs at 40 cpm—high frequency (HF)—and 10 cpm—low frequency (LF). Control sensitivities were optimized for each combination, and learning curves were obtained from separate groups of inexperienced Ss. With the HF input the differentiating control system gave a smaller mean error than the integrating system ($p < .01$), whereas with the LF input it gave the larger mean error ($p < .001$). With both inputs the amplifying system tied for 1st place. The reasons for this were elucidated from oscillographic charts.

Most previous work on tracking has used positional or higher order control systems, such as the integrating (rate control) system of Figure 1B. No one appears seriously to have considered orders of control lower than positional. The present experiment examined the differentiating control system of Figure 1A, using random inputs and a pursuit display. If the input had been given a fixed value, the task would have been comparable to one used by Lincoln (1954) whose subjects had to wind a handwheel at a constant rate in order to keep the indicator of a tachometer type of display on a specified mark.

METHOD

Subjects. These were enlisted men in the British Royal Navy aged between 17 and 31 years. None of them had had much experience of tracking. Their scores on Intelligence Test AH4 (Heim, 1955) were found to be unrelated to their tracking performance.

¹ The apparatus was designed and built by A. Davidson from funds made available by the British Army Personnel Research Committee. The subjects were supplied by the Royal Navy. Financial support from the British Medical Research Council is also gratefully acknowledged.

Apparatus. The display consisted in a double-beam cathode ray tube (CRT), diameter 6.0 inches, which had a thin black horizontal line dividing its face into upper and lower halves. The input was represented by a spot of light, diameter .075 inch, which moved in a vertical dimension. The amplitude of its movement was adjusted so that as far as possible it traversed the full height of the face of the CRT without disappearing off the edge. A bright horizontal line $.9 \times .05$ inch had to be kept superimposed upon the spot.

The control was a hollow aluminium rod 4.0 inches long, designed to be held between the thumb and forefinger of the supported hand. It was mounted directly below the face of the CRT, and moved 110 degrees on either side of its center point in the same direction as the controlled line moved on

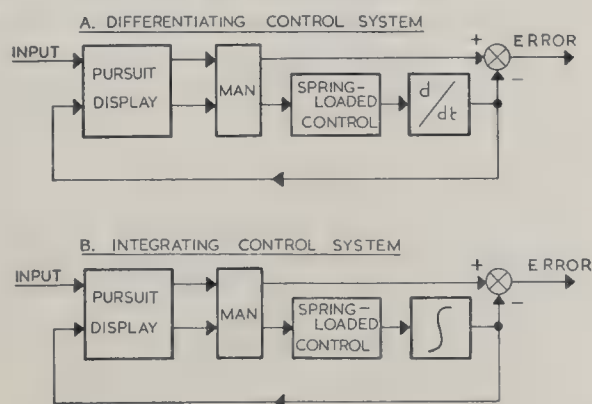


FIG. 1. Comparison of differentiating and integrating control systems.

TABLE 1
CONDITIONS AND MEAN ERROR AFTER PRACTICE

Control system	Control move- ment for 1-inch display movement	N	Summed mean error	
			M	SE
HF input			Average of Sessions 6 and 7	
Differentiating	24° per second	6	60.0	2.2
Amplifying	Low sensitivity 13°	6	54.1	2.8
	Higher sensitivity 4°	8	60.4	3.6
Integrating	Low sensitivity 20° for 1.0 second	6	71.3	2.8
	Higher sensitivity 2° for 1.0 second	9	81.0	2.1
LF input			Average of Sessions 5 and 6	
Differentiating	14° per second	6	41.3	1.3
Amplifying	40°	6	8.2	.3
Integrating	Low sensitivity 20° for 1.0 second	6	9.6	.7
	Higher sensitivity 5° for 1.0 second	5	16.1	2.5

the CRT. It was spring centered, and required an approximately constant force of about 2.0 ounces to move it away from center, about .5 ounce of which was combined stiction and friction. With the control centered, the controlled line on the face of the CRT was stationary when using the integrating control system and lay directly beneath the central thin black line when using the amplifying system. With the differentiating system the controlled line assumed the central position whenever the control came to rest.

The error voltage passed through a switching device which converted negative voltages to positive and was then fed to a summing device. A two-pen Evershed and Vignoles oscillograph could be used to record simultaneous samples of any two voltage functions.

Input frequencies and control sensitivities. The basic input consisted in white noise covering about four octaves, having equal energy per cycle, and an upper cutoff of about 20 decibels per octave. Two frequencies were used: a high frequency (HF) input with the upper cutoff set at 40 cycles per minute, and which was otherwise unfiltered; and a low frequency (LF) input with the cutoff set at 10 cycles per minute and which then passed through two independent low-pass R-C networks with time constants of 3.0 and .6 seconds.

For each frequency and control system the optimal control sensitivities were determined experimentally on separate groups of five subjects. Each subject tracked with a different one of five representative sensitivities each day, according to a Latin square design. The subjects were drawn from the same population as the experimental subjects, but did not perform in the experiment proper. The differentiating control system proved to have a relatively narrow region of optimal control sensitivity; the sensitivities on either side giving statistically reliable increases in error were in the ratio

of about 1:7. With the LF input the same was true for the amplifying control system. However with the HF input the region of optimal control sensitivity was rather wider; a change by a factor of 15 was found to be insufficient to produce a statistically reliable increase in error at the sensitive end of the region.

The integrating control system had an even wider region of optimal control sensitivity; a change by a factor of 60 did not produce a statistically reliable increase in error, although sensitivities much greater than this could not be tested, since some subjects could not hold the line on the face of the CRT at the start of the experiment. In the three instances in which a statistically reliable limit could not be ascertained at the sensitive end of the region of optimal control sensitivity, two values were selected for the main experiment, one towards the lower end, the other nearer the probable middle of the region. The values are listed in Table 1.

Experimental design. A random group design was used to avoid transfer effects between conditions. All the subjects available at any one time were allocated to the same condition to encourage competition between them. Figure 2 shows, for each control system, the number of sessions which were spread over the 2 weeks for which the subjects were available. A session consisted in five periods of tracking of just over 1.0 minute each.

Procedure. At the start of the first session the subject was allowed to manipulate the control stick for about 15 seconds without any input. The aim of each session was to maximize the rate of learning. The experimenter always watched the subject, and after each period of tracking told him how to improve his performance, in addition to giving him a score representing his mean error, and comparing it with the previous performance of himself and his colleagues. At the end of each day the scores of all the subjects in the group were

entered on a chart fixed outside the door of the experimental room. Before starting a session the experimenter always drew the subject's attention to his scores on the chart and compared them with those of his colleagues. Oscillographic recordings were made immediately after the last period of tracking in the last session.

Scoring and calculations. The mean error was summed over the last 60 seconds of each period of tracking. In Figure 2 the mean has been expressed as a percentage of the average error which accrued when the controlled line on the face of the CRT remained stationary in its central position, in order to cancel out random session-to-session variations in the amplitude of the input.

Sample measurements were taken from the oscillographic charts along the general lines described in Poulton (1962). In addition, a "wobble" score was computed as follows from the records with the HF input. The number of times that the error was exactly zero was counted and from this was subtracted twice the number of input cycles with an amplitude of 5% or more of the full range of movement of the input shown on the record. Since a wobble was taken to be a double movement out and back, the remainder was divided by two. The score gives an indication of the amount of very high frequency in the response.

In comparing the differences between means, two-tailed *t* tests have always been used except where stated.

RESULTS

The learning curves are given in Figure 2. Table 1 compares the average error in the sixth and seventh sessions with the HF input and the average in the fifth and sixth sessions with the LF input. It shows that in the three instances where two sensitivities of control were used, the less sensitive always proved to be the better ($p < .05$). With the HF input the differentiating control system was not different from the amplifying system at either of the two sensitivities tested ($p > .05$). The integrating control system was significantly the worst, even with the better sensitivity ($p < .05$). In contrast, with the LF input the differentiating control system was by far the worst ($p < .001$), while with the better sensitivity the integrating system was no different from the amplifying system.

The oscillographic charts showed that with the HF input the integrating control system produced a smaller mean amplitude of response than did the amplifying control system ($p < .05$, one-tailed test), while the amplifying control system produced the largest mean

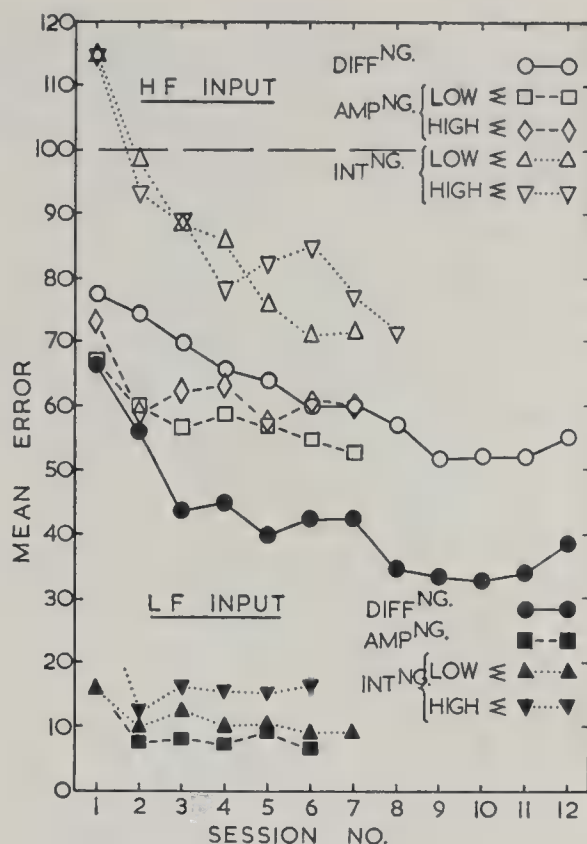


FIG. 2. Learning curves comparing the control systems, sensitivities (Σ), and input frequencies. (Each function represents the mean error of a group of from five to nine subjects. Points show the average of five 1-minute periods of tracking by each subject, adjusted to make 100 the level if the subject had not responded.)

time lag ($p < .05$). The differentiating control system gave a mean score of 36 wobbles per minute; the amplifying system, 13; and the integrating system, 4 ($p < .05$). With the LF input, only the differentiating control system showed any appreciable error on the oscillographic charts. The mean amplitude of response was about one third of the input, while the very high frequency components in the response made up the difference intermittently.

DISCUSSION

The smaller time lag with the HF input when tracking with the *differentiating control system* as compared with the amplifying system was presumably due to the elimination of the subject's movement time. In order to change the position of the controlled line on the CRT, the subject did not have to change

the position of his control; he had only to alter its rate of movement. This advantage with the HF input offset the disadvantage of amplifying the very high frequencies in the response (reflected in the high wobble score), and left the differentiating system with a mean error no larger than that produced with the amplifying control system. This was presumably because even a small reduction in time lag was a great advantage with the HF input; while the very high frequencies in the response function did not affect the mean error very much since there was still an appreciable time lag, and the amplitude of the high frequencies was small relative to the amplitude of the input. It follows that where it is extremely important to minimize the subject's time lag, but an exaggerated tremor can be tolerated, it is possible that a differentiating control system could be designed which would be preferable to any alternative system.

With the LF input, the small reduction in time lag effected by the differentiating control system was a relatively minor advantage, while the amplified very high frequencies in the responses function were a major source of error. The amplified high frequencies probably also at least partly concealed from the subject the small mean amplitude of his low frequency movements matching the input.

The *integrating control system* introduced an additional time lag into the response function while the rate put on by the control movement changed the position of the controlled line on the CRT. The oscillographic charts with the HF input showed that the subject was able to compensate for this and actually had a smaller mean time lag than with the amplifying system, although at the cost of a reduced amplitude of movement of the response function. The reversals of the recorded control movements of the best subject synchronized on average with the changes in the direction of movement of the input. This involved predicting the input ahead for the duration of his reaction time and thus running the risk of an inaccurate prediction (Poulton, 1957, Table 1), although the smaller amplitude of movement of the response function somewhat reduced the cost

of the risk. With the HF input, the reduced amplitude of the response function, by which the subject compensated for the disadvantage of the additional time lag inserted by the control system, was nowhere near offset by the relatively minor advantage of the attenuation of the very high frequencies in the response function, which was reflected in the very small wobble score. In contrast with the LF input, the additional time lag was a relatively minor disadvantage which could be overcome without any reduction in the amplitude of the response function. It was thus offset by the now major advantage of the attenuation of the very high frequencies. The relative positions of the functions in Figure 2 for the integrating and amplifying control systems with the HF and LF inputs are in line with the results of Chernikoff and Taylor (1957, Figure 1 and Table 2).

A difficulty met with both the differentiating and integrating control systems, which does not appear to have been pointed out previously, resulted from the 90 degrees of phase advance or retard. In two quarters of each input cycle the subject had to move his control in the same direction as the input was moving; let us call this a compatible control-display relationship. But in the other two quarters he had to move his control in the reverse direction to the direction of movement of the input, an incompatible relationship. The rapid succession of compatible and incompatible relationships was not easy to learn, although the oscillographic charts showed that it had always been mastered by the end of the experiment.

REFERENCES

- CHERNIKOFF, R., & TAYLOR, F. V. Effects of course frequency and aided time constant on pursuit and compensating tracking. *J. exp. Psychol.*, 1957, **53**, 285-292.
- HEIM, A. W. *Manual for the Group Test of General Intelligence AH4*. London: National Foundation for Educational Research, 1955.
- LINCOLN, R. S. Rate accuracy in handwheel cranking. *J. appl. Psychol.*, 1954, **38**, 195-201.
- POULTON, E. C. On prediction in skilled movements. *Psychol. Bull.*, 1957, **54**, 467-478.
- POULTON, E. C. On simple methods of scoring tracking error. *Psychol. Bull.*, 1962, **59**, 320-328.

(Received September 5, 1962)

FACTORS AFFECTING PERCEPTUAL INTEGRATION OF ILLUSTRATED MATERIAL¹

JAMES M. MCKENDRY, MONROE B. SNYDER, AND STEPHEN GATES²

HRB-Singer, Incorporated, State College, Pennsylvania

This paper is concerned with perceptual integration, i.e., the process by which parts of an illustration are coded into perceptual units. Interest was centered about the stability of the process and some measurable variables which might influence it. 20 Ss were shown a series of 58 illustrations. A moderate degree (approximately 54%) of interindividual consistency was shown. Test-retest revealed a high degree (approximately 94%) of intra-individual consistency. 4 variables (space separation, border separation, object similarity, and object interdependence) bore a significant relation to the particular perceptual integration process studied. Practical implications of the results are discussed.

This paper is concerned with one particular type of human perceptual activity and the way it might influence subjects' reactions to advertising illustrations. In Miller's (1956) terms, it involves the coding of "bits" of information (Shannon & Weaver, 1949) into "chunks" of information upon which the subject later acts, i.e., apparently, subjects do *not* passively transmit each physical element of displayed information but group (code) the elements (or bits) in some way and attach labels to these groupings (Osgood, 1957). If this particular type of activity, termed *perceptual integration*, can be unambiguously described within an advertising context, it is possible that a new measure can be derived to describe an illustration, viz., how many "perceptual units" are contained. This variable may bear some interesting relationships to advertising "effectiveness."

The research described within had two major objectives: to determine if a proposed set of measurable variables influenced perceptual integration and to determine the nature and extent of their influence. *Perceptual integration* was defined as the means by which *elements* of advertising illustrations become transformed into perceptual units, an

"illustrational element" being an environmental object (Gibson, 1959), i.e., a self-contained bounded mass which is distinguishable from other masses by form and edge cues. Environmental objects studied were chairs, books, people, animals, packages, and cars. A "perceptual unit" was defined as a cognitive category, composed of one or more illustrational elements, forming the subject's coded representation of the contained elements.

Factors Affecting Integration

Wertheimer (1923), using dots as illustrational elements, found that individuals grouped dots into "figures" or "units" on the basis of the following: nearness or proximity in the field of view, sameness or similarity, common fate or similarity in directional orientation, good continuation or good figure, conformity with the individual's momentary set or *Einstellung*, and past experience or custom. Only two of these variables appear to be directly measurable, i.e., physical proximity and element similarity. Based upon a review of studies summarized by Woodworth and Schlosberg (1954) a third measurable variable, contour, was added. It was defined as "an abrupt change gradient in either brightness or color" (p. 412), or, mathematically, as the second differential of brightness. One final variable was added to take account of one type of interaction between environmental objects. Working definitions of each of the four measurable variables

¹ This research was supported by the Advertising Research Foundation, Incorporated, by a contract to HRB-Singer, Incorporated. The views expressed are those of the authors. Acknowledgment is made to Paul M. Hurst and Robert E. Stover for their helpful suggestions and criticisms.

² Now with the Mitre Corporation.

are provided below, phrased in a form allowing their use by raters.

Space Separation. While this variable (the counterpart of proximity) is continuous, its more important features can be maintained by using five categories, each of which represents an order of magnitude: "high overlap" which includes those cases where the areas of any two objects overlap by 50% or more, "slight overlap" which includes cases of mutual object overlap between 6% and 49%, "adjacent placement" which includes cases starting with the juxtaposition of objects up to and including 5% overlap, "slight separation" which begins with those cases where objects are separated by 1% of the mean object width³ and ends with the case where separation is equal to the mean object width, and "significant separation" which includes all those cases beyond the upper limit of the previous category.

Border Separation. While there are numerous possible categories within the limits of this scale, which is a correlate of contour, only three occur with any frequency in illustrations: cases where objects are contained within a common border; cases where objects, while not being contained in the same border, have no border separating them; and cases where objects are completely separated by a border. The presence of border is defined as the use of lines or edges, other than object edges, or the use of a sudden "significant" change in color or brightness, e.g., going to white space from a blue background or other ground discontinuities. Less abrupt changes, such as a man standing in front of a building, while a woman nearby is outlined against the sky, which are part of many typical scenes, are not included in the ground "change" rule. When two perceptual elements are contained within a common border and yet have a second border separating them, the more proximal one is assumed to be critical, i.e., the second border.

Object Similarity. There are three order-of-magnitude categories used for this variable:

³ Computation of the mean object width can be complicated by the fact that a single object can have numerous widths. In such a case, it is suggested that the largest width measure be used for each object.

instances of very high similarity, where the *same* object appears twice, e.g., "before and after" type illustrations; instances where *similar* objects appear, e.g., two human beings, two books, two geometric figures, two chairs, etc.; and instances of little similarity, where two *different* objects appear, e.g., a table and a chair, a man and a dog, a cigarette and a lighter, a car and a man, etc.

Object Interdependence. The first category includes cases of object *interdependence*. Here, there is a strong *functional* connection between two perceptual elements, e.g., a cigarette being lit by a lighter, a table holding up some books, two people talking to each other, a man sitting in a chair, a girl riding a horse, etc. The second category which is determined by a process of elimination includes cases where object interdependence does not occur.⁴

Hypotheses

The following hypotheses were advanced:

Hypothesis 1. Individuals are consistent (both within themselves and among themselves) in the way in which they integrate illustrational elements into perceptual units.

Hypothesis 2. The probability that two specific illustrational elements will be integrated into a single perceptual unit is an increasing function of object interdependence and of the degree of object similarity. It is a decreasing function of the use of border separation and of the use of space separation.

METHOD

Scope

Experimental materials were selected to cover the range of variation provided by the four classification factors. Since these variables had 5, 3, 3, and 2 categories, respectively, a set of 90 possibilities exists where each possibility describes a potential relationship existing between two illustrational elements.

However, some of these 90 possibilities describe events which occur rarely in advertising, e.g., it

⁴ A possible third category could evolve from a subdivision of the interdependence category into instances of slight and strong interdependence (one possible basis of differentiation being whether or not both objects are included in the interdependence, or if only *one* of the elements is interacting, e.g., one person talking to another person who is obviously not listening).

TABLE 1
ILLUSTRATIONAL EVENTS USED AS STIMULI

Border Separation × Object Similarity		Space Separation × Object Interdependence									
Border separation	Object similarity	High overlap		Slight overlap		Adjacent		Slight		Significant	
		I	NI	I	NI	I	NI	I	NI	I	NI
Within same border	Same		5		15	21	26		36		50
	Similar	1	6	11	16	22	27	32	37	45	51
	Different	2	7	12	17	23	28	33	38	46	52
No border separation	Same		8		18		292		39		53
	Similar	3	9	13	19	24	30	39	40	47	54
	Different	4	10	14	20	25	31	35	41	48	55
Complete border separation	Same							42			56
	Similar							43			57
	Different							44		49	58

Note.—Numbers indicate cards while blank areas indicate events not used since they were unusual illustrations. Letters I and NI indicate cases of interdependence and no interdependence, respectively.

is unusual to have complete border separation coupled with object overlap since superimposition is necessitated. In fact, only 58 of the 90 possibilities were found by reviewing over 500 magazine advertisements. Table 1 shows all 90 possibilities. The 58 cases which were found to occur are numbered consecutively in the table.

Stimuli

Each of the 58 "plausible" relationships between illustrational elements shown in Table 1 was represented by a separate stimulus. Thus, each stimulus served as an example of two particular illustrational elements having some specific relationship to each other: Stimulus Number 1 demonstrated the first number relationship described in Table 1 (two similar, interdependent, elements—bounded by a common border—placed in such a way that the elements had a high overlap), Stimulus Number 12 demonstrated the twelfth numbered relationship described in Table 1 (two different, yet interdependent, elements—bounded by a common border—placed in such a way that the elements had a slight overlap), etc.

Stimuli were cut out portions of actual magazine illustrations pasted on 10 \times 12 inch cards. The following rules governed the construction of all stimulus cards: two and only two elements were placed on each card; each cutout was centered; size perspectives of the two elements on a card were constant; and factors such as the use of color, glossiness, and element size perspective were allowed to vary between cards (stimuli) but not within cards (among the elements in a single stimulus). A wide range of illustrational elements were employed, e.g., people, books, cars, bottles, packages, cans, candies, tires, animals, and glasses. Examples of stimuli

depicting the two relationships described above were as follows: Stimulus Number 1 (Card 1) was a colored cutout of two full glasses of beer placed within a common border so that one was slightly behind the other, thereby allowing a high degree of element (glass) overlap; Stimulus Number 5 (Card 5) was a colored cutout showing a full martini glass held up "to dry" by a clothes pin in such a way that a slight overlap occurred between the two elements (the glass and the clothes pin)—a border encased both elements.

Instructions

A measure of the tendency to group any two perceptual elements into a single perceptual unit was operationally defined as a verbal response of "one" to the instructions given below.

We are going to show you a number of cards with different advertisements. Each advertisement is composed of a number of elements. We want to know if the elements appear to form one or two units.

This is not a test. There are no right or wrong answers. We are primarily interested in your first impression; therefore, you will be given only one second to view a card (ad).

Are there any questions?

If subjects asked questions, the instructions were paraphrased again; no additional explanation was added since it was felt to be extremely likely that biases or sets would result from such a procedure.

Subjects

A sample of 20 subjects, composed of 14 males and 6 females, in the 20–30 year age range, was

chosen from the personnel at HRB-Singer, Incorporated. The sample, which contained secretaries, service personnel, and some engineers, none of whom were under the administrative control of the experimenters, was not random. Subjects had no knowledge of the purpose of the study.

Procedure

The set of 58 cards was randomized before each presentation. As the subject responded, an experimenter sitting across a table from the subject, sorted the cards into piles of "one" or "two" responses. Exposure times were approximately 1 second in duration. Once the sequence was completed, a second experimenter recorded the responses while the subject was performing another task.

This second task involved the use of 45 more stimulus cards which were not part of this study. While the subject was responding to this interposed task, the illustrational cards were again randomized for a second presentation, which began after the interposed task was completed. The two responses, by the same subject, to each card were later compared to judge individual consistency.

In order to eliminate distortions in the individual consistency figure caused by immediate memory effects, a second retest was run 3 months after the initial test. In this comparison, one of the two initial responses—either the test or the retest—was randomly selected to be compared with a new set of retest responses for a group of 10 subjects randomly selected from the original 20.

Analysis

Hypothesis 1, regarding consistency *within* an individual over time, was tested by examining two measures. The first was obtained by comparing the responses made by each subject during the initial test to those made by the same subjects during the initial retest. The second was obtained by comparing the responses made by each subject during the initial test (or initial retest) with those made by the same subjects 3 months later. In both instances, each subject received a score indicating the number of times (out of 58 possibilities) that he gave the same response to each of the 58 stimuli.

A second aspect of Hypothesis 1 was also tested, i.e., in regard to the consistency that the *entire sample* of 20 subjects showed in responding to the 58 stimulus cards. This was done by examining the division of 20 responses for each stimulus card into "one" or "two" categories and determining the probability of obtaining such a division if the two probabilities were .50 in both cases. Binomial tests were used in these analyses (Siegel, 1956).

Hypothesis 2 was tested by four separate analyses of variance, each of which was a three-variable design, i.e., Treatment (orders of magnitude along each variable) \times Subjects \times Matched Advertisements design. Matched advertisements were those classified the same with respect to three variables

but which differed on the fourth, e.g., for the case of object interdependence Cards 1 and 6, Cards 2 and 7, Cards 3 and 9, etc., constituted the set of matched pairs of advertisements (see Table 1).

Thus, the same responses, grouped in different ways, were employed in all four F tests. Each cell entry in the four three-dimensional matrices was the response of a particular subject to a particular stimulus card, i.e., either a 1 or a 2. Both matched advertisements and treatments were treated as "fixed constants," and the Subject \times Treatments interaction used as the error term for the main effects (McNemar, 1956).

The techniques employed in testing Hypothesis 2 have two disadvantages. First, tests for classification factors are not independent of each other, i.e., certain cells are used in more than one analysis. This is comparable to using multiple measures on the same subjects. Second, the range of values possible within a particular cell (one subject's response to one card) was restricted to one of two numerical values resulting in a binomial distribution of scores for a cell. Such departures from normality would result in a departure from the theoretical F distribution.

The first effect discussed above was at least partially controlled since the exact groupings of individual cells used in each analysis differed considerably. The second does not appear to be as serious as one might imagine, largely as a consequence of the large number of degrees of freedom in each computation. The F tests were mainly concerned with arithmetic means of collections of cells (e.g., an overall mean for a treatment of the means for all subjects); by the central limit theorem one can expect the distribution of means to approach normality. Inspection of the data verified this expectation. However, as an added precaution, nonparametric analyses were conducted to verify the significance level of the effects of each classification factor; in the case of the interdependence factor, the Wilcoxon test was used (Siegel, 1956); the Friedman test was used in all other analyses (Siegel, 1956).

RESULTS AND CONCLUSIONS

Regarding Hypothesis 1

Analysis of the initial test-initial retest data showed that the 20 subjects gave the same answers in 91% of the cases, a figure which is considerably higher than chance agreement ($p < .001$ by a t test).⁵ Analysis of the initial

⁵ In order to take account of response bias caused by marginal discrepancies, e.g., a tendency by a particular subject to call most cards a "one," the two t tests were computed as follows: A four-fold (χ^2 type) classification was made between (1A) test and (1B) retest responses and (2A) one-unit and (2B) two-unit responses, thereby allowing estimates of "expected" agreements to be computed. These were then compared to the obtained number of

test-later retest data for the 10 subjects showed a mean agreement of 81% ($p < .001$). These data indicate that subjects are reasonably consistent over time in the way in which they group, or do not group, elements into perceptual units.

Computations of binomial probabilities of obtaining observed splits between one- and two-unit responses for the 20 subjects showed that 43 of the 58 stimulus cards were categorized with $p < .10$, 31 of the 58 cards were categorized with $p < .05$, and 25 of the 58 stimuli were categorized with $p < .01$. Since some extreme splits between one- and two-unit responses can be expected on the basis of chance alone, a second test was conducted to determine how many stimuli (cards) of the 58 stimuli would receive a split having a $p < .10$ simply as a consequence of performing so many binomial tests. The Poisson distribution was used as an approximation to the binomial method suggested by Wilkinson (1951) using p values of .10, .05, and .01, respectively. Each test showed that obtaining the observed number of extreme splits simply as a consequence of performing 58 different tests was unlikely ($p < .0001$). Thus, it appears that there is a reasonable amount of agreement among subjects concerning the way in which particular stimuli were interpreted.

Regarding Hypothesis 2

Tables 2 through 5 show the results of the four analyses of variance; each of the classification variables showed a significant main effect. These findings were confirmed by the nonparametric tests: the space separation factor had a $\chi^2 = 39.54$ with $df = 4$ ($p < .001$); the border factor had a $\chi^2 = 29.005$ with $df = 3$ ($p < .001$); the object similarity factor had a $\chi^2 = 9.70$ with $df = 3$ ($p < .01$); and the object interdependence factor had a $T = 6$ with an $N = 20$ ($p < .005$). In all cases a significant first-order interaction between subjects and treatments was ob-

agreements to obtain a difference score for each subject. These data yielded smaller t values than those obtained by assuming that the expected probability of agreement per card per subject was .50. However, all t values were easily significant ($p < .001$).

TABLE 2

RESULTS OF ANALYSIS OF THE EFFECTS OF
SPACE SEPARATION OF GROUPING
TENDENCIES

Source	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>
Space separation (SS)	4	15.25	3.812	28.879***
Advertisements (A)	9	28.64	3.182	
Subjects (S)	19	49.21	2.590	
SS \times S	76	10.03	.132	1.451*
SS \times A	36	18.85	.524	5.758***
A \times S	171	39.14	.229	2.516***
SS \times S \times A	684	62.47	.091	

* $p < .05$.

*** $p < .001$.

served, indicating that the subjects did not respond uniformly to the treatments.

An interaction having more serious implications was noted in the analyses for space separation and border effects. In these cases, the first-order interaction between treatments and advertisements was significant, indicating that the effects of these classification factors were partially confounded with the characteristics of particular advertisements selected, i.e., particular cards chosen. In the "border analysis," this second type of interaction was not only statistically significant, but also sizeable in the proportion of the total variance associated with it (i.e., the size of the mean square value), although it was still much less than the variation associated with border effect. However, in the "space separation analysis," this interaction did not account for a very large proportion of variance.

The information contributed by analyses of the first order interactions and the main

TABLE 3

RESULTS OF ANALYSIS OF THE EFFECTS OF THE USE OF
BORDER ON GROUPING TENDENCIES

Source	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>
Use of border (B)	2	7.70	3.850	4.880*
Advertisements (A)	16	9.70	1.616	
Subjects (S)	19	19.20	1.010	
B \times S	38	3.00	.789	6.743***
B \times A	12	13.40	1.117	9.547***
S \times A	114	19.70	.173	1.479***
B \times S \times A	456	26.6	.117	

* $p < .05$.

*** $p < .001$.

TABLE 4

RESULTS OF ANALYSIS OF THE EFFECTS OF OBJECT SIMILARITY ON GROUPING TENDENCIES

Source	df	SS	MS	F
Object similarity (O)	2	5.20	2.600	8.075**
Advertisements (A)	12	44.37	3.698	
Subjects (S)	19	37.70	1.984	
O × S	38	12.23	.322	3.157***
O × A	24	14.90	.621	6.088***
S × A	228	30.60	.134	1.313
O × S × A	456	46.67	.102	

** $p < .01$.
*** $p < .001$.

TABLE 5

RESULTS OF ANALYSIS OF THE EFFECTS OF OBJECT INTERDEPENDENCE ON GROUPING TENDENCIES

Source	df	SS	MS	F
Object interdependence (I)	1	12.53	12.53	18.646***
Advertisements (A)	21	41.83	1.99	
Subjects (S)	19	35.60	1.87	
I × S	19	12.77	.672	5.695***
I × A	399	47.74	.120	1.017
S × A	21	4.06	.193	1.636
I × S × A	399	47.15	.118	

*** $p < .001$.

effects indicates the following: there are individual differences which should be studied further, features of particular advertisements can cause changes in grouping behavior (the second point indicates that it might be profitable to add further classification factors), and it appears that the four classification variables investigated are of importance since they account for the largest proportion of variance in three of the four analyses conducted.

A description of the nature of each of the four treatment effects is shown in Figure 1 where the mean number of units per card is plotted as a function of increasing degrees of magnitude for each of the four scales. The "categories" shown in the figure are ordered from "low to high" as follows: for space separation, each successive category indicates an *increase* in space separation; for the border scale, each successive category indicates a *higher degree* of border separation;

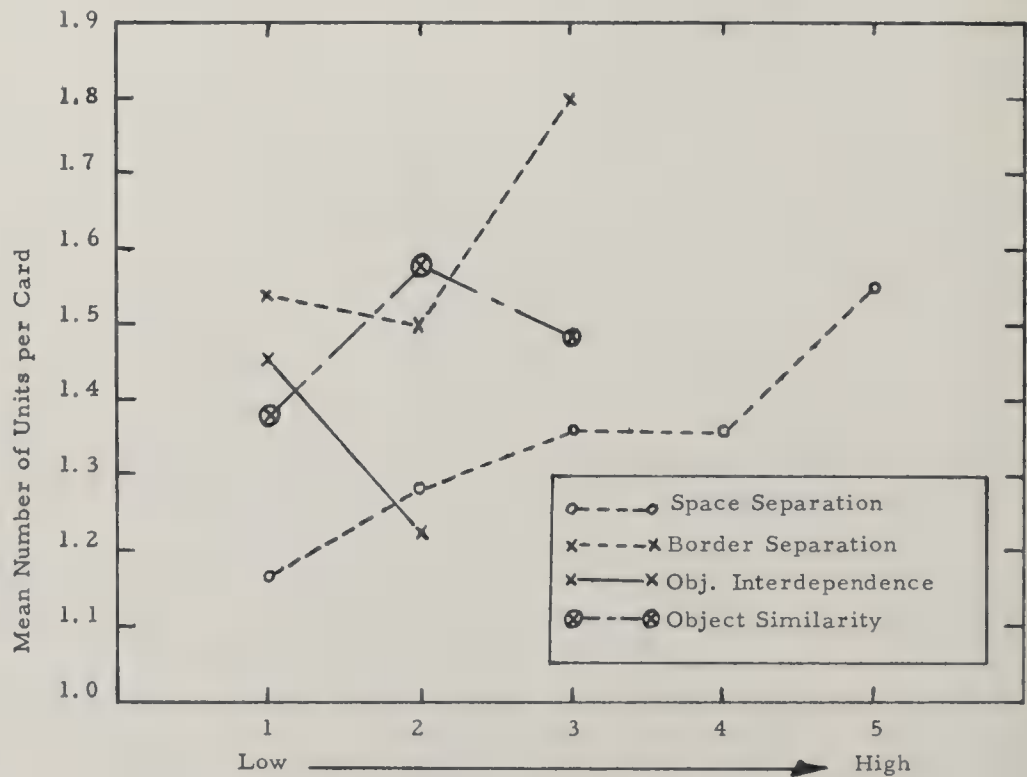


FIG. 1. Illustration of the effects of each of the four classification factors on the mean number of units reported.

for the object similarity scale, each successive category indicates an *increase* in object similarity; and for the interdependence scale, the first category indicates no interdependence while the second indicates interdependence. The figures shows that the a priori hypotheses were substantiated in all cases *except* for object similarity where a substitution of similar objects for the same objects induces the anticipated result while a substitution of different objects for similar objects yields results not anticipated, i.e., an *increase* in grouping tendency. There does not appear to be any simple explanation of this unexpected outcome.

IMPLICATIONS

The obtained data provide reasonable support for Hypotheses 1 and 2, with the exception of the object similarity hypothesis. If the description of the effects of this variable are changed, it becomes possible to derive an empirically supported tentative rating scheme by which assessments can be made of the number of perceptual units of illustration contained in an advertisement. After evolving such a scheme, described by Snyder (1961), it was found that two raters scoring a total of 20 ads taken from *Printed Advertising Rating Methods* (Advertising Research Foundation, 1956) took 10–90 seconds per ad and that their assessments had a product-moment correlation of .899. Snyder (1961) also found that perceptual unit scores had significant correlations with three criteria of advertising effectiveness. It would therefore appear that the results have some practical consequences.

However, before applying findings evolved from the study described within, some precautions should be kept in mind. First, it should be remembered that advertisement exposure times of from 1 to 2 seconds were used. If much shorter exposure times occur, it is doubtful that a subject would be able to discriminate between "same" or "similar" things. Conversely, if very long exposure

times are used, finer discriminations might begin to appear. The time of 1 second was used since it is approximately the amount of time that a subject would spend if he were hastily scanning an advertisement.

Second, the effect of size perspective is important. Pilot studies conducted by the authors indicate that use of two elements having different size perspectives invariably leads to a "two" response. That is, the elements are *not* fused, but are kept distinct, with each one becoming a perceptual unit.

REFERENCES

- GIBSON, J. J. Perception as a function of stimulation. In S. Koch (Ed.), *Psychology: A study of a science*. Vol. 1. *Sensory, perceptual, and physiological formulations*. New York: McGraw-Hill, 1959. Pp. 136–162.
- MCNEMAR, Q. *Psychological statistics*. (2nd ed.) New York: Wiley, 1955.
- MILLER, G. A. The magical number seven, plus or minus two, or some limits in our capacity for processing information. *Psychol. Rev.*, 1956, **63**, 81–97.
- OSGOOD, C. E. A behavioristic analysis of perception and language as cognitive phenomena. In, *Contemporary approaches to cognition*. Cambridge: Harvard Univer. Press, 1957.
- ADVERTISING RESEARCH FOUNDATION. *Printed advertising rating methods*. New York: ARF, 1956.
- SHANNON, C., & WEAVER, W. *The mathematical theory of communication*. Urbana: Univer. Illinois Press, 1949.
- SIEGEL, S. *Nonparametric statistics in the behavioral sciences*. New York: McGraw-Hill, 1956.
- SNYDER, M. B. The measurement and control of visual display efficiency. Report No. 238-F, 1961, HRB-Singer, Incorporated, State College, Pennsylvania.
- WERTHEIMER, M. Untersuchungen zur lehre von der Gestalt. II. *Psychol. Forsch.*, 1923, **4**, 301–350. Cited by R. A. Woodworth and H. Schlosberg, *Experimental psychology*. (Rev. ed.) New York: Holt, 1954. Pp. 408–409.
- WILKINSON, B. A statistical consideration in psychological research. *Psychol. Bull.*, 1951, **48**, 156–158.
- WOODWORTH, R. S., & SCHLOSBERG, H. *Experimental psychology*. (Rev. ed.) New York: Holt, 1954.

(Received June 20, 1962)

SOME PERSONALITY AND BEHAVIORAL FACTORS RELATED TO BIRTH ORDER ¹

EWART E. SMITH AND JACQUELINE D. GOODCHILDS

Serendipity Associates, Los Angeles, California

A study was conducted using 165 firemen in large and small firehouses to test the hypothesis that first borns learn to interact more successfully because of their stronger dependency and affiliation needs. It was also predicted that this greater interactional skill would become more apparent the more complex the social situation. As predicted, first borns had less self-confidence. They also conformed more, were more efficient problem solvers in a group situation, and were more often the official leader of their work group. However, these group behaviors were only related to birth order in the larger and more complex groups.

Schachter's (1959) synthesis of a number of findings on the relationships between birth order and various behaviors, and his linkage of this research with Festinger's (1954) theory of social comparison processes have provided us with a heuristic new theory. The best statement of it can be taken from Schachter (1959):

It will be recalled that the diverse data on the effects of birth order have all been interpreted in terms of a common notion—dependence or the degree to which the individual relies on others as sources of approval, support, help, and reference. . . . Designating this dimension of reliance on others as dependence, it should be anticipated that first-born and only persons would place more reliance on social means of evaluation than would later-born persons. . . . When placed in a situation some aspect of which requires evaluation, early-born individuals are more likely than later-born persons to seek out others as a means of evaluation; when together with others in such a situation, early-born are more likely than later-born individuals to rely on others in evaluating their own opinions and emotional states [pp. 131-132].

From these theoretical statements we can postulate that first borns place greater reliance on interaction with others, as a means of solving their problems, than they do on their own actions. We would therefore expect to find first borns conforming more in groups, and would also anticipate that first

borns would have lower scores on measures of self-reliance or self-confidence.

The greater affiliative tendencies, in time of need, of first borns should lead to more interactions with others, followed by the feedback inherent in most social interaction situations. Consequently, one would hypothesize that first borns would be more interested in and experienced with social interaction, and therefore should be more successful at such interaction. In addition, the hypothesized greater interactional skills of first borns should be more apparent the larger and more complex the social situation, due to the greater number and complexity of the interactions.

These considerations led us to the following hypotheses:

1. First-born subjects will be lower than later borns on a personality measure of self-confidence.
2. First-born subjects will be higher than later borns on:
 - a. Conformity in their group.
 - b. Formal rank in their group.
 - c. Task efficiency in a group situation under time pressure.
 - d. Perceived clarity of the social structure of their group.
3. The hypothesized relationships between these social behavior and attainment measures and birth order will be stronger in larger, more complex groups than in smaller and simpler groups.

¹ This work was supported by the Behavioral Sciences Division, Air Force Office of Scientific Research of the Office of Aerospace Research, under Contract No. AF 49(638)-1000.

METHOD

Subjects

The subjects were 165 Los Angeles firemen. Ten small single company and 10 large double company firehouses were used. Each firehouse was visited once and data obtained on the entire crew present. The single company houses had one captain and averaged 5 men per crew and the double company houses had two captains and averaged 11 men per crew. Firemen can and do transfer from one firehouse to another and it cannot be assumed that any differences between the men in the single and double houses are random.

Procedure

In an introductory statement the research was explained as an attempt to find out how men work together in groups, in order to help the government in planning crews for such special situations as space vehicles and submarines.

The subjects first completed a questionnaire and then worked on two problems, giving a preanswer to each problem just before the group discussion and individual postdiscussion answers afterwards.

The first group task was to discuss the following question for 15 minutes and to reach a group decision:

You are to be marooned on a desert island with no immediate hope of rescue. You are allowed to take with you three and only three things. What will you choose to take?

This problem was deliberately ambiguous to provide a good discussion.

The second group problem was to discuss and give an answer within 8 minutes to the Maier (1952) Horse Trading Problem:

A man bought a horse for \$60 and sold it for \$70. Then he bought it back for \$80 and again sold it for \$90. How much money does he make in the horse business?

Measures

Self-Confidence. This was measured by the *K* scale. High scores on the *K* scale from the MMPI (Meehl & Hathaway, 1946) are obtained by denying that various negative statements (e.g., "Criticism or scolding hurts me terribly") are applicable to one's self. Consequently, the *K* scale is a measure of a subject's own statement of his mental health, and was used as a measure of self-confidence, or what might more technically be termed the strength of ego defenses (see Smith, 1959; Sweetland & Quay, 1953; Wheeler, Little, & Lehner, 1951).

Conformity. Conformity scores were obtained by comparing individual postdiscussion answers to the Desert Island task with the answers given by the group. These scores could range from zero to three according to how many of the three group choices the subjects utilized.

TABLE 1

MEAN CONFORMITY AND ROLE CLARITY BY
BIRTH ORDER AND GROUP TYPE

Type	<i>N</i>	Mean conformity	Mean role clarity
Large groups			
First born	41	2.54	94.34
Later born	63	2.18	88.38
<i>t</i>		2.33*	2.30*
Small groups			
First born	23	2.56	92.35
Later born	26	2.31	93.42
<i>t</i>		1.10	0.26

* $p < .05$.

Perceived Clarity of Group Role Structure. This was measured by an 18-item Likert-type Role Clarity scale designed to indicate how clear the group role structure appeared to each subject. Typical items are: "We know what to expect of each other in this group," "I understand the people in this group," and (scoring reversed) "It's hard to guess how we'd act in an emergency."

Formal Rank. Information as to whether each subject was an ordinary fireman or a fire captain was obtained.

Task Efficiency in a Group Situation. This measure was each individual's post-group-discussion answer to the Horse Problem, scored correct or incorrect. Individual prediscussion answers were also obtained.

RESULTS ²

The data on the behavioral measures were analyzed separately for large and small groups. Only-born subjects were always classified as first borns.

On the *K* scale the 70 first borns had a mean of 16.49 compared to 17.98 for the 95 later borns, with a *t* of 2.45 ($p < .05$).

The results on the conformity measure were consistent with previous research (Schachter, 1959) with first borns conforming more than later borns in large groups, as shown in Table 1, with a *t* of 2.33 ($p < .05$). However, the differences in the small groups, while in the same direction, were not significant.

² The *N* is 153 on some measures, as 12 subjects, 8 from the large groups and 4 from the small, were not present for the two group tasks. All statistical tests are two-tailed.

TABLE 2

STATUS AND TASK EFFICIENCY BY
BIRTH ORDER AND GROUP TYPE

Type	Formal rank		Horse trading problem	
	Cap- tain	Fire- men	Cor- rect	Incor- rect
Large groups				
First born	12	34	32	9
Later born	7	59	36	27
χ^2	4.61*		4.80*	
Small groups				
First born	4	20	18	5
Later born	5	24	17	9

* $p < .05$.

The first borns also scored higher on the Role Clarity scale, in the large groups, with a t of 2.30 ($p < .05$). There were no significant differences in the small groups, as can be seen in Table 1.

First borns were more frequently captains in the large groups than were later borns, as shown in Table 2, with a chi square of 4.61 ($p < .05$). There was no difference in the small groups.

The first borns also scored higher on the Horse Problem, as shown in Table 2, but again the difference was only significant in the large groups ($\chi^2 = 4.80$, $p < .05$). It is interesting to note that the first borns were not superior to the later borns on the Horse Problem prior to the group discussion, indicating that the first borns benefited more from the group discussion than did the others.³

DISCUSSION

Our findings appear to support and extend the theorizing of Schachter (1959). Of par-

ticular importance are the data on the K scale, in view of the previous lack of birth order differences on personality tests. Our interpretation of the K scale (on normals), as a measure of self-confidence, and the lower K scale scores of first borns supports Schachter's (1959) and Sears' (1950) description of first borns as more dependent.

The first born's reduced self-confidence and hence greater dependency on others apparently produces the greater conformity seen in our data as well as Schachter's (1959). The results on the measure of perceived clarity of the group role structure are another indication of this greater dependency on and interest in interpersonal relations.

The superiority of the first borns in the large groups on the Horse Problem, which has a correct answer, is particularly illuminating in the light of the results on the conformity measure on the ambiguous Desert Island task. The first borns were not correct more frequently than the others on the Horse Problem before the group discussion, and were only more superior in those groups which solved the problem. And it was only in these latter groups that the first borns conformed to the group solution. These results indicate that first borns not only seek out and respond to others more frequently, particularly under stress, as demonstrated by Schachter (1959), but they also seem to profit more from the association. Conformity tendencies are not so strong where there is a reality factor that first borns disregard correctness. Hence we should not think of the apparent dependency of first borns in a negative manner, although dependency has such a connotation in our culture. This seeking out of others for help seems to be functional. And it may have been this responsiveness to the thoughts and feelings of others that resulted in more first borns attaining the rank of captain in the large, complex firehouses. This finding may be related to Schachter's (1959, p. 78) observation that first-born students belonged to more organizations than did later borns. However, it must be noted that we do not know what mediated birth order and attaining formal leadership in the large groups. Perhaps further exploration of

³ A more parsimonious explanation of the Horse Trading Problem data might be that this is conformity behavior, similar to that on the more ambiguous Desert Island Problem, as most of the groups, 14 out of 20, gave the correct answer to the Horse Problem. However, separate analyses of the data on the correct and incorrect groups revealed that the first borns conformed significantly to the group answers in the former but not in the latter.

the psychology of the first born will lead to successful attacks on the problem of leadership, a problem swept under the rug years ago.

It must be noted that the behavioral relationships found held only in the large, more complex firehouses. This strengthens our theoretical assumptions that the relationships between birth order and the behaviors and attainments investigated are mediated by a greater responsiveness to others and ability to interact successfully, and also suggests future areas of research, such as family size, for increasing our understanding of birth order psychology.

REFERENCES

- FESTINGER, L. A theory of social comparison processes. *Hum. Relat.*, 1954, 7, 117-140.
- MAIER, N. R. F., & SOLEM, A. R. The contribution of a discussion leader to the quality of group thinking: The effective use of minority opinions. *Hum. Relat.*, 1952, 5, 277-288.
- MEEHL, P. E., & HATHAWAY, S. R. The *K* factor as a suppressor variable on the Minnesota Multiphasic Personality Inventory. *J. appl. Psychol.*, 1946, 30, 525-564.
- SCHACHTER, S. *The psychology of affiliation*. Stanford: Stanford Univer. Press, 1959.
- SEARS, R. R. Ordinal position in the family as a psychological variable. *Amer. sociol. Rev.*, 1950, 15, 397-401.
- SMITH, E. E. Defensiveness, insight, and the *K* scale. *J. consult. Psychol.*, 1959, 23, 275-277.
- SWEETLAND, A., & QUAY, H. A note on the *K* scale of the Minnesota Multiphasic Personality Inventory. *J. consult. Psychol.*, 1953, 17, 314-316.
- WHEELER, W. M., LITTLE, K. B., & LEHNER, G. F. J. The internal structure of the MMPI. *J. consult. Psychol.*, 1951, 15, 134-141.

(Received August 23, 1962)

INDIVIDUAL DEFFERENCES IN SELECTION DECISIONS¹

PATRICIA M. ROWE

McGill University

The cognitive variable "category width" was examined for its applicability to the problem of individual differences in selection decisions. Accept or reject decisions for 100 "applicant" descriptions were made by 146 Ss. Analysis of the decisions showed striking between-individual differences and within-individual consistencies in the number of applicants accepted. Differences in the width of the category "acceptable applicants" were found to be related to past learning and present motivational state. Evidence that category width is a general trait was also found. It was concluded that much decision variance can be accounted for in terms of the category width of the interviewer.

It is a well-documented observation that when several personnel interviewers separately assess the same applicant they arrive at very dissimilar decisions (see review of studies by Wagner, 1949). One conclusion frequently drawn from this finding is that personnel selection decisions based on interviews are extremely unreliable and, consequently, that selection procedures should be drastically modified. However, the observed individual differences among personnel interviewers should be of intrinsic interest to psychologists, and their study might eventually make it possible to control interview unreliability at its source—in the interviewer himself. In the closely related area of categorization of simple, inanimate objects, investigators have taken the approach of regarding unreliability as an individual difference problem. These investigators have been successful in isolating a number of "response styles" (McGee, 1962) or "cognitive controls" (Gardner, Holzman, Klein, Linton, & Spence, 1959) underlying individual differences in categorization or judgment.

One of these response styles has been termed "category width" by Bruner and his co-workers (e.g., Bruner, Goodnow, & Austin,

1956). A typical category-width study is that by Bruner and Tajfel (1961). Subjects first viewed a slide containing a cluster of 20 dots. Subsequently, they were presented with slides containing clusters of 20-30 dots and asked to judge for each stimulus whether or not it contained 20 dots. Category width was defined by the number of stimuli judged "20 dots," and subjects were classified as broad or narrow categorizers depending on whether they made many or few such judgments. In another paper Bruner (1957) suggested that the accessibility of a category (and thus category width) is determined by two factors—the learned probabilities of occurrence of events in the individual's world, and the search requirements that are dictated by his need states.

If the principle of category width applies to decisions about persons, it may be that much of the "unreliability" of selection decisions could be accounted for in terms of individual differences in the width of the category "acceptable applicants" among the interviewers. The present investigation was designed to test this notion, and, further, to examine the role of Bruner's two factors of past learning and present motivational state in determining category width.

METHOD

Subjects. Because previous studies (Webster, 1959) indicate that selection standards vary as a function of company employment practices and job requirements, it was necessary to use as subjects interviewers employed by the same company and regularly selecting employees for the same position. The most readily available group meeting these

¹ This paper is based on parts of a PhD thesis submitted to McGill University. The research was supported by Defence Research Board of Canada Grants, No. 9435-53 to E. C. Webster and No. 9425-10 to D. Bindra. The assistance of members of the Canadian Army Personnel Selection Service is gratefully acknowledged. The author is indebted to E. C. Webster, D. Bindra, and M. P. Bryden for their advice and assistance.

requirements was the Personnel Selection Service of the Canadian Army. Of the 263 Personnel Selection Officers (PSOs) first approached for this study, 146 completed all tests and constituted the sample of subjects.

Two characteristics of the Canadian Army should be mentioned here because of their relevance to the analysis of the data: (a) All Canadian servicemen are volunteers; in recent years there have been many more applications than openings. Consequently, the PSO must *select* some applicants and reject many more on other than medical grounds. (b) The PSOs in this study are of two types: Regular Force PSOs are those employed in the Army on a full-time basis; Militia PSOs are roughly equivalent to members of the Active Reserve in the United States, and serve in the Army for approximately one-half day per week.

Test Materials. Pilot studies had indicated that applicants themselves change as a function of being interviewed several times; therefore, written descriptions of applicants were chosen as stimuli. One hundred fictitious "applicant" descriptions were constructed from 60 characteristics, 30 favorable and 30 unfavorable, all chosen from Sydiaha's (1958) Applicant Description Checklist. These characteristics had been found by him to be significantly related to "accept" or "reject" decisions made by eight PSOs. Each applicant description was composed of 6 of the 60 characteristics. Three of the characteristics were favorable, the other three unfavorable, arranged in random order. By using an equal number of favorable and unfavorable characteristics, individual differences were probably exaggerated, but on the other hand, the likelihood of obtaining spurious agreement among the subjects was reduced. The following is a sample item:

This applicant has no commitments: he is completely independent. He seems cocky, a bit of a smart aleck, ready to tell anybody off. He tends to be an active participant in outdoor activities. The applicant makes a good impression; his appearance, language, all round ability and bearing make him stand out in a group. He has often left jobs following disagreements with superiors. The applicant seems to be an argumentative person.

All descriptions were presented to the subjects as "applicants for the Regular Force, Canadian Army, who had met minimum Infantry standards on all criteria." The task for the subject was to accept or reject each applicant.

In addition to the 100 descriptions, the Characteristic Rating Scale, a slightly modified form of Sydiaha's Applicant Description Checklist, was included in the test material. The subjects rated each of 119 characteristics, which included the 60 used in constructing the applicants, on a 7-point scale from 1 (very unfavorable) to 7 (very favorable). Thus the more favorably a characteristic was rated, the higher was its score.

Procedure. All test materials were mailed to the subjects, who returned the completed forms to the

investigator. In order to get a better estimate of the reliability of the decisions, the applicant descriptions were mailed in three groups (consisting of 35, 35, and 30 descriptions, respectively). The Characteristic Rating Scale was mailed separately after the judgments had been received.

RESULTS

The mean number of applicants (descriptions) accepted by the 146 subjects was 34.86 out of 100, with a standard deviation of 20.02, and with a range from 0 to 92 applicants accepted. An analysis of variance, in which the variation between subjects was compared to the variation within subjects from one mailing of the test to another, yielded an F of 17.00 ($p < .01$), and thus demonstrated significant individual differences in the number of applicants accepted. Moreover, the correlations between the numbers of applicants accepted in the three mailings of the descriptions showed a high degree of within-individual consistency. The correlation between Parts 1 and 2 was .849; between Parts 2 and 3, .863; and between Parts 1 and 3, .828. It may be concluded, therefore, that there are both significant between-individual differences and within-individual consistencies in the width of the category acceptable applicants.

Not only do the subjects vary in the number of applicants accepted, but also, the applicants themselves differ in their likelihood of being accepted. The mean number of acceptances received by the 100 applicants was 50.89 out of 146, with a standard deviation of 26.51, and with a range of 8–116 acceptances received. Because of these differences in applicant acceptability we may raise the question of whether the differences in the category widths of the subjects are non-random—that is, do the subjects, regardless of the number of applicants they accept, tend to select those applicants most frequently accepted by the group as a whole? Such an analysis is equivalent to determining the scalability of the subjects.

In the present study subject scalability was determined with Bryden's (1960) rank-biserial correlation. In all, 142 coefficients were calculated (in four cases the statistic was not applicable) of which the mean coefficient was .569, the standard deviations was

TABLE 1
ANALYSIS OF NUMBER OF APPLICANT DESCRIPTIONS
ACCEPTED BY KIND OF SERVICE AND RANK

Group	<i>N</i>	<i>M</i>	<i>SD</i>	<i>t</i> or <i>F</i>
Regular Force				
Majors	8	27.38	4.18	<i>F</i> = .06
Captains	28	28.78	20.04	
Lieutenants	5	31.00	10.22	
Militia				
Majors	31	26.29	15.29	<i>F</i> = 7.09**
Captains	58	41.83	20.33	
Lieutenants	16	41.81	21.38	
Regular Force	41	28.76	17.95	<i>t</i> = 2.33*
Militia	105	37.24	20.28	
Total Group	146	34.86	20.02	

* *p* < .05.
** *p* < .01.

.141, and the range was from .179 to .873. Of these coefficients, 132 were significant beyond the .01 level, 7 were significant at the .05 level, and 3 were not significant. It is clear that the subjects in this study are scalar—almost all PSOs, regardless of the number of applicants they accepted, tended to select those applicant descriptions most acceptable to the group as a whole.

Is the dimension of acceptability of the applicants related to any other properties of the applicants? The mean rating of each characteristic making up the applicant descriptions was determined from the Characteristic Rating Scale. These scores were then combined appropriately to yield two scores for each applicant description: the mean rating assigned the three favorable characteristics, and the mean rating assigned the three unfavorable characteristics. The correlations of the number of acceptances received with these two scores were .471 and .616, respectively; with the total of the two scores, .713; the multiple correlation was .801. Correlations of such magnitude indicate that the more favorably an applicant's characteristics were rated the more frequently was he accepted by the subjects.

Now let us turn to a consideration of the sources of individual differences in category width. Table 1 presents the findings of an

analysis of number of applicants accepted as a function of kind of service (Regular Force or Militia) and rank (Major, Captain, or Lieutenant). The difference between the Regular Force and Militia PSOs was determined by a *t* test; differences between ranks were tested by one-way analyses of variance and the means compared two at a time. Regular Force PSOs accepted fewer applicants than did Militia PSOs, thus reflecting the different standards of acceptance in the ordinary duties of the two kinds of officers. Moreover, Militia Majors accepted fewer applicants than did Captains or Lieutenants, which suggests that with increasing experience in personnel selection the interviewer learns that not all men make good soldiers and consequently reduces his number of acceptances.

The number of applicants accepted was also found to be related to the manner in which the subjects responded on the rating scale. For purposes of the present study the Characteristic Rating Scale consists of four types of characteristics: favorable (*N* = 30) and unfavorable (*N* = 30) characteristics used in constructing the applicants, and favorable (*N* = 26) and unfavorable (*N* = 33) characteristics not used in the descriptions. For each subject the mean rating assigned to each of these four subgroups of characteristics was computed. (The higher the mean the more favorably were the characteristics of any type rated.) The correlations between the number of applicants accepted and the mean ratings of the characteristics used in the applicant descriptions were .308 (favorable) and .608 (unfavorable); for the characteristics not used in the descriptions they were .348 (favorable) and .495 (unfavorable); all correlations are significant beyond the .01 level. These correlations show that both favorable and unfavorable characteristics are rated more favorably by subjects who accepted many applicants than by those who accepted few.

DISCUSSION

To the extent that between-individual differences and within-individual consistencies in the width of the category acceptable ap

plicants have been demonstrated, the present investigation has shown that the concept of category width is applicable to decisions in personnel selection. Although written descriptions were used as applicants in this study, this conclusion is at least warranted in the area of selection decisions (e.g., those concerning applicants for graduate studies) made on the basis of letters of recommendation to which the present descriptions bear much resemblance. It is not unreasonable to expect that similar, large individual differences may be found in other decision-making areas. For example, we might find that dentists vary in their judgments about "enough decay to necessitate extraction of the tooth," that judges differ regarding "seriousness of a particular crime," and that baseball umpires disagree on decisions of "within the strike zone."

Interviewers differ in the number of applicants they accept, but agree as to the relative acceptability of the various applicant descriptions. That is, regardless of the number of applicants a subject may accept, he tends to select those who are most acceptable to the group as a whole. It follows from this finding that in most cases a fairly accurate prediction of whether or not a particular subject will accept a particular applicant can be made from the knowledge of two variables: the rank order of that applicant in the whole group, and the category width of that subject. Moreover, applicant rank order is not necessarily an ad hoc measure, but may be determined from how favorable his characteristics are. The importance of applicant rank order in decision making is not surprising; what the present study contributes is the fact that, because the interviewers are scalar, category width also plays a significant role in accounting for differences associated with selection decisions.

Several characteristics of the subjects were found to be indicative of the sources of individual differences in category width. Kind of service, rank, and ratings assigned to applicant characteristics were all related to the number of applicants accepted. Differences as a function of kind of service and rank are evidence for the importance of Bruner's (1957) two factors of motivational states and past learning in

determining category width. The finding that subjects show closely related differences in their standards of acceptance of both applicants and single characteristics is further support for the notion that category width is a general trait displayed in a variety of decision situations, as suggested by Fillenbaum (1959).

To summarize, then, this study has emphasized certain characteristics of the interviewer in producing individual differences in selection decisions. Apparently, applicants can be meaningfully ordered along a dimension which is defined by how favorable their characteristics are. Whether a particular applicant will be accepted is a joint function of his position on this dimension and the category width of the interviewer he sees. Future investigations should be directed towards the relation between individual differences in category width and such variables as the accuracy of decisions, prescribed numbers of applicants to be accepted, and personality and motivational characteristics of the interviewer.

REFERENCES

- BRUNER, J. S. On perceptual readiness. *Psychol. Rev.*, 1957, **64**, 123-152.
- BRUNER, J. S., GOODNOW, JACQUELINE, & AUSTIN, G. A. *A study of thinking*. New York: Wiley, 1956.
- BRUNER, J. S., & TAJFEL, H. Cognitive risk and environmental change. *J. abnorm. soc. Psychol.*, 1961, **62**, 231-241.
- BRYDEN, M. P. A non-parametric method of item and test scaling. *Educ. psychol. Measmt.*, 1960, **20**, 311-315.
- FILLENBAUM, S. Some stylistic aspects of categorizing behavior. *J. Pers.*, 1959, **27**, 187-195.
- GARDNER, R. W., HOLZMAN, P. S., KLEIN, G. S., LINTON, HARRIET B., & SPENCE, D. P. Cognitive control: A study of individual consistencies in cognitive behavior. *Psychol. Issues*, 1959, **1** (Whole No. 4).
- MCGEE, R. K. Response style as a personality variable: By what criterion? *Psychol. Bull.*, 1962, **59**, 284-295.
- SYDIAHA, D. The relation between actuarial and descriptive methods in personnel appraisal. Unpublished doctoral thesis, McGill University, 1958.
- WAGNER, R. F. The employment interview: A critical summary. *Personnel Psychol.*, 1949, **2**, 17-46.
- WEBSTER, E. C. Decision making in the employment interview. *Personnel Admin.*, 1959, **22**, 15-22.

(Received August 27, 1962)

A FACTOR ANALYSIS OF EXPERIMENTAL SOCIAL DESIRABILITY AND RESPONSE SET SCALES¹

ALLEN L. EDWARDS

University of Washington

Intercorrelations between 19 response set scales, based upon the scores of 110 students, were factor analyzed and the factors rotated orthogonally. Edwards' SD scale and 6 experimental social desirability scales had their highest loadings on the 1st factor. 3 scales containing neutral items in which the probability of a True response to the items varied between scales were found to have their highest loadings on 3 different factors. Scales designed to measure the tendency to give deviant True responses to items with socially undesirable scale values, to items with socially desirable scale values, and to items with neutral scale values were found to have their highest loadings on 3 different factors. The tendency to give deviant True responses to items with socially undesirable scale values was found to be related to the tendency to give deviant False responses to items with socially desirable scale values. The tendency to mark items as doubtful and the tendency to answer items marked doubtful as True were identified as 2 factors unrelated to social desirability tendencies.

Considerable research interest has developed in response sets as they may be reflected in responses to items in personality scales.² The two response sets which have been investigated most thoroughly are acquiescence, as described by Cronbach (1946, 1950), and social desirability, as described by Edwards (1957). Scores on personality scales have been regarded as being susceptible to the influence of acquiescent tendencies to the degree to which there is an imbalance in the True-False keying of the items in the scale (Couch & Keniston, 1960; Fricke, 1956; Jackson & Messick, 1958, 1961; Messick & Jackson, 1961; Wiggins, 1959) and to the influence of social desirability tendencies to the degree to which there is an imbalance in the social desirability keying of the items in the scale (Edwards, 1957, 1961; Edwards & Diers, 1962; Edwards, Heathers, & Fordyce, 1960).³

¹ This study was supported in part by Research Grant M-4075 from the National Institute of Mental Health, United States Public Health Service. Carol J. Diers, Sharon Feeney, and James A. Walsh, my research assistants, participated in one way or another in the data collection, the scoring of the test records, and in the statistical analyses.

² General reviews of the literature may be found in Cronbach (1946, 1950), Edwards (1957), Jackson and Messick (1958), McGee (1962), and Wiggins (1962).

³ Edwards and Walsh (1963) have recently reported evidence to show that the intensity of the

Edwards and Heathers (1962), Edwards and Diers, and Walker (1962), and Edwards and Diers (1962) have shown that first factor loadings of Minnesota Multiphasic Personality Inventory (MMPI) scales are highly correlated with the proportion of items in the scales keyed for socially desirable response and this finding has been regarded as supporting the interpretation of the first factor of the MMPI as a social desirability factor. It is also known that first factor loadings of MMPI scales are correlated with the proportion of items keyed True in the scale (Edwards & Diers, 1962; Edwards et al., 1962; Messick & Jackson, 1961) and this result can be viewed as supporting the interpretation of the first MMPI factor as an acquiescence factor. Edwards and Diers (1962) have shown, however, that if the MMPI is administered under conditions where acquiescent tendencies may be assumed to be minimized and social desirability tendencies maximized, the correlation between the proportion of items keyed True and first factor loadings does not disappear as it should if acquiescent tendencies are not operating, but rather that the correlation remains essentially the same as that found

social desirability keying of items in a scale is also an important variable which is related to the degree to which scores may be influenced by social desirability tendencies.

when the MMPI is administered under standard instructions. It seems clear on the basis of other evidence presented by Edwards and Diers that their subjects were not responding to the MMPI items in terms of acquiescent tendencies. Thus, they conclude that the correlation between the proportion of items keyed True in MMPI scales and the first factor loadings must be at least in part the result of the confounding of the True keying with the social desirability keying in the MMPI scales.

In addition to social desirability and acquiescent interpretations of the first factor of the MMPI, Barnes (1956a) presents evidence to show that MMPI scales which have high loadings on the first factor are correlated with the tendency to give deviant True responses, whereas scales which have substantial loadings on the second factor are correlated with the tendency to give deviant False responses. It has been shown by Edwards (1953), however, that the probability of a True response to a personality item is a linear increasing function of the social desirability scale value of the item and there is evidence to indicate that a similar relationship holds for MMPI items (Hanley, 1956). This means, of course, that, in general, the modal response to items with socially desirable scale values is True and that the modal response to items with socially undesirable scale values is False. Thus items keyed for deviant True responses would, in general, also tend to be keyed for socially undesirable True responses and items keyed for deviant False responses would also tend to be keyed for socially undesirable False responses.

High scores on a deviant True scale may thus measure the tendency to give deviant True responses the tendency to acquiesce, the tendency to give socially undesirable responses to items with socially undesirable scale values, or some combination of these tendencies. Similarly, high scores on a deviant False scale may measure the tendency to give deviant False responses, the tendency to dissent, the tendency to give socially undesirable responses to items with socially desirable scale values, or some combination of these tendencies. In research with the MMPI, using

standard MMPI scales, these various response sets all tend to be confounded with one another and with the scoring keys of the scales. To take but one example, Welsh's (1952) A or first factor scale contains 39 items of which 38 are keyed True; 37 of the 39 items are keyed for socially undesirable responses, and 38 of the 39 items are keyed for deviant True responses. Thus when the A scale is found to have a high loading on a factor in common with various other MMPI scales possessing correlated or confounded item properties similar to those found in the A scale, it is little wonder that this factor can be interpreted by some investigators as a deviant True factor, by other investigators as an acquiescence factor, and by still other investigators as a social desirability factor.

Because of the confounding in the deviant True keying, the social desirability keying, and the True keying, because of the similarities and dissimilarities in item keying when items appear in more than one MMPI scale and thus produce either spuriously high or spuriously low correlations between scales, and because of the pathological content of many MMPI items, it seems important to investigate further the relationship between factor loadings of personality scales and response sets with scales with nonoverlapping items and in which the item content relates to more normal aspects of personality. By designing new experimental scales in which items are included on the basis of their psychometric properties, rather than relying upon existing empirical scales in which items are included according to their ability to differentiate between criterion groups or on rational scales in which items are selected in terms of their content, it may be possible to obtain some differentiation between the response sets which have been called social desirability, acquiescence, and deviance.

METHOD

Social desirability ratings of 2,824 experimental personality items were obtained from a group of 47 male students regularly enrolled at the University of Washington. Each student was paid a standard fee for attending five rating sessions. At each session the students rated 600 items on a 9-point social desirability continuum, following the instructions

given by Edwards (1957). For each item the mean rating on the social desirability continuum was obtained and this value is referred to as the social desirability scale value of the item. A serial sample of 176 items rated during the first four sessions was repeated at the last session. The test-retest reliability of the scale values of these 176 items was .97.

Another independent group of 110 male students served as subjects in the experiment. Each student was paid a standard fee for attending two testing sessions each week for a period of 3 weeks. At each of the first five sessions subjects were given two test booklets consisting of 300 experimental items each. They were not given the second booklet until they had completed the first. The task assigned to the subjects was to read carefully each item and to determine whether or not they believed the item accurately described them. It was emphasized in both oral and written instructions that if they were in doubt as to the correct response, i.e., if they were not sure as to whether the item did or did not describe them, they were to put an X between the True and False columns corresponding to the item on their IBM answer sheet. They were also instructed to give their best guess on each doubtful item by marking the item True or False. At the last session, the subjects were administered a battery of standard personality scales and inventories under standard instructions.

Fourteen scales believed to measure types of response sets were derived from responses to the experimental personality items. It will be convenient to refer to the set of items upon which each of the 14 scales is based as a scale. It is to be emphasized, however, that these are not personality scales in the usual sense of the word. In determining which items were to be included in a given scale, no attention was paid to the content of the items and, presumably, as far as content is concerned, the items within a given scale are heterogeneous. In addition to the 14 experimental scales, scores were obtained on five of the standard scales administered during the last testing session. A brief description of the 19 scales is given below.

Scales 1 and 2 are experimental scales consisting of items with socially undesirable scale values keyed for socially undesirable (True) responses. The item probabilities of endorsement fall along the regression line of the probability of a True response on social desirability scale value. Since the modal response to these items is False, the items are also keyed for what Barnes (1956b) has called deviant True responses.

Scales 3 and 4 are also experimental scales consisting of items with socially undesirable scale values keyed for socially undesirable (True) responses. In both scales the item probabilities of endorsement exceed predicted probabilities, based upon the regression line of probability of endorsement on social desirability scale value, by at least one standard error. These two scales are similar to Scales 1 and 2 except that the items have somewhat

larger probabilities of eliciting socially undesirable responses. Since the modal response is False, the items are also keyed for deviant True responses.

Scales 5, 6, and 7 are experimental scales in which all of the items have scale values in the neutral interval, 4.5 to 5.5, on the 9-point social desirability continuum. Items in all three scales are keyed True. Scale 5 is a neutral scale in which all of the items have relatively low probabilities of being answered True. High scores on this scale may be indicative of either an acquiescent tendency or of a tendency to give deviant True responses or both. In Scale 6 the item probabilities of endorsement are between .47 and .53. The items are thus of a kind which Wiggins (1962) has described as being of high controversiality. Presumably high scores on Scale 6 may be indicative of the tendency to acquiesce, but not of the tendency to give deviant True responses. In Scale 7 the item probabilities of a True response are all equal to or greater than .58. High scores on this scale may thus measure the tendency to acquiesce or the tendency to give modal or common responses to neutral items. Low scores on the scale may be indicative of dissentience or of the tendency to give what Barnes (1956b) has called deviant False responses.

Scales 8 and 9 are experimental scales containing items with socially desirable scale values keyed for socially desirable (True) responses. The item probabilities of endorsement fall along the regression line of the probability of a True response on social desirability scale value. Presumably, high scores on these two scales should measure the tendency to give socially desirable (True) responses to items with socially desirable scale values. The tendency to give socially desirable responses is regarded by Edwards (1957, 1961) as the bipolar opposite of the tendency to give socially undesirable responses. Thus, according to social desirability considerations, these two scales should have loadings of an opposite sign on a common factor on which Scales 1-4 also have loadings. Since the items in Scales 8 and 9 and Scales 1-4 are all keyed True, acquiescence considerations would predict loadings of the same sign of these scales on a common factor. Since the modal response to the items in Scales 8 and 9 is True, low scores on these scales may also be regarded as measuring the tendency to give deviant False responses. Both Barnes (1956b) and Wiggins (1962) have reported that on MMPI items, the tendency to give deviant False responses is uncorrelated with the tendency to give deviant True responses. Thus, according to the deviation hypothesis, Scales 8 and 9 should not have loadings on any factor in common with Scales 1-4, if either set of scales is relatively factorially pure. If the two sets of scales, Scales 1-4 and Scales 8 and 9, are factorially complex, then it is possible that both sets of scales could have loadings on two or more common factors and still be uncorrelated.

Scale 10 is an experimental scale containing items with socially desirable scale values keyed for socially desirable (True) responses. The scale differs from

Scales 8 and 9 in that the modal or common response to the items is not the socially desirable (True) response. All of the items have a probability of endorsement equal to or less than .45. High scores on this scale should thus measure the tendency to give socially desirable but relatively uncommon True responses. High scores may also be indicative of the tendency to acquiesce or of the tendency to give deviant True responses, since the modal response is False.⁴

Scale 11 consists of the 51 items in the experimental set of 2,824 items which had the largest probabilities of being marked doubtful. Upon examination of the social desirability scale values of these items it was found that 3 items had scale values between 4.0 and 5.0 and that all of the remaining 48 items had scale values on the socially desirable side of the neutral point. All items in the scale are keyed for True responses.

Scale 12 is Edwards' (1957) Social Desirability (*SD*) scale. The scale has 9 items with socially desirable scale values keyed True and 30 items with socially undesirable scale values keyed False. Thus, all the items are keyed for socially desirable responses but the scale differs from the other experimental social desirability scales in that it has some degree of balance in its True-False keying.

Scale 13 is the Crowne and Marlow (1960) Social Desirability (*M-C SD*) scale. The scale contains 18 items with socially desirable scale values keyed True and 15 items with socially undesirable scale values keyed False. All items are keyed for socially desirable responses. The scale differs from the other social desirability scales included in the analysis in that it is modeled after the Lie (*L*) scale of the MMPI. The *L* scale was designed to measure the tendency of subjects to give improbable but socially desirable responses (Hathaway & McKinley, 1951).

Scale 14 is Wiggins' (1959) *Sd* scale. The scale contains 31 items with socially desirable scale values keyed True and 9 items with socially undesirable scale values keyed False. The items contained in the scale were selected on the basis of their ability to differentiate between subjects given the MMPI under standard instructions and under instructions to falsify their responses in order to look good.

Scales 15 and 16 are based upon Couch and Keniston (1960) items. A substantial number of the items in these two scales are not typical of those found in personality scales but rather resemble those found in attitude and opinion questionnaires. These items relate to relatively broad and general beliefs and values. Typical of these items are such statements as:

There are days when one awakes from sleep without a care in the world, full of zest and eagerness for what lies ahead.

⁴ An attempt was made to develop a scale containing items with socially undesirable scale values which could be keyed for deviant False responses. However, a sufficient number of items of this kind could not be found in the experimental pool of 2,824 items.

Happiness is one of the primary goals of life.

The world is teeming with opportunities and promises of success for anyone with sufficient imagination to perceive them.

The vast majority of men are truthful and dependable.

Scale 15 consists of 35 items keyed True. The items in Scale 15 are those which Couch and Keniston found to have a mean agreement rating on a 7-point Disagree-Agree rating scale between 5.0 and 5.9 and they refer to Scale 15 as a High Mean (HM) scale. Another set of 40 items was selected from the list of Couch and Keniston items to make up Scale 16. The items in Scale 16 differ from those in Scale 15 in that the mean agreement rating on the 7-point Disagree-Agree rating scale for the items in Scale 16 is between 3.8 and 4.2. It will be convenient to refer to Scale 16 as a Middle Mean (MM) scale.

Scale 17 is the total number of doubtful or ? responses which a subject gave to the first 2,400 experimental personality items.

Scale 18 is the conditional probability of a True response given that an item has been marked doubtful by a subject. Scores on this scale were obtained by finding the number of True responses which each subject subsequently gave to those items he had marked doubtful (score on Scale 17). The number of True responses to these items was then divided by the total number of doubtful responses to give the subject's score on Scale 18.

Scale 19 is based upon the number of identical (True-True plus False-False) responses to the 176 duplicated experimental items. Scores on this scale presumably measure the tendency to respond consistently to personality items when social desirability tendencies and acquiescent tendencies are both free to operate. It seems reasonable to believe that if either or both of these tendencies are operating upon both the first and second occasions when an item is presented, then consistency of response will be increased.

Scores on the 19 scales were intercorrelated and factor analyzed by the method of principal components. Ten factors were extracted. The first principal factor accounted for 33.1% of the total variance and the subsequent factors accounted for 17.4, 8.5, 5.9, 5.3, 4.6, 4.1, 3.4, 3.1, and 2.8% of the total variance, respectively. The 10 factors together accounted for 88.0% of the total variance. The 10 factors were rotated orthogonally using a varimax IBM 709 program.⁵

RESULTS

The 19 scales are listed in Table 1 along with a brief description of each scale to serve

⁵ The principal axis factor analysis and rotation programs for the IBM Type 709 are described in mimeographed reports by F. Nussbaum, D. A. Marshall, and P. Roosen-Runge (1961) and G. Burket (1961).

TABLE 1
MEAN PROBABILITIES OF A KEYED RESPONSE, $P(K)$, MEANS, STANDARD DEVIATIONS, AND
KUDER-RICHARDSON FORMULA 21 COEFFICIENTS FOR 19 PERSONALITY SCALES

Scales and keying	Scale values	$P(K)$	\bar{X}	s	K-R 21
1. SUD: ^a 75 True	2.5-3.0	.11	8.57	9.69	.93
2. SUD: 67 True	3.0-3.5	.19	12.51	10.47	.92
3. SUD: 23 True	2.0-3.0	.26	6.04	4.28	.79
4. SUD: 54 True	3.0-4.0	.45	24.30	9.85	.88
5. Low Neutral: 50 True	4.5-5.5	.25	12.36	4.35	.52
6. Neutral: 46 True	4.5-5.5	.49	22.62	5.69	.66
7. High Neutral: 50 True	4.5-5.5	.65	32.60	6.04	.70
8. SD: 75 True	7.0-7.5	.80	59.95	9.19	.87
9. SD: 51 True	7.5-8.0	.91	46.45	4.75	.83
10. SD Deviant: 55 True	≥ 6.0	.36	19.76	7.47	.80
11. SD Doubtful: 51 True	4.0-8.0	.65	33.30	7.42	.81
12. Edwards' SD: 9 True, 30 False		.83	32.39	5.20	.82
13. M-C SD: 18 True, 15 False		.44	14.39	5.62	.77
14. Wiggins' Sd: 31 True, 9 False		.36	13.09	3.82	.41
15. Couch-Keniston HM: 35 True		.81	28.49	4.31	.74
16. Couch-Keniston MM: 40 True		.52	20.73	4.70	.56
17. Total number ? responses		.16	393.73	308.19	.93 ^b
18. $P(T/?)$ on Scale 17			.60	.14	.73 ^b
19. Consistency: 176 repeated items		.85	149.19	9.42	.75

^a SUD = socially undesirable.
^b These two coefficients are split-half coefficients. The two scores for each subject are based upon blocks of 1,200 items each.

as a convenient reminder of the nature of the scale. The mean probability of a keyed response to the items in each scale is given in the column headed $P(K)$. The last column in the table gives the Kuder-Richardson Formula 21 (K-R 21) lower bound estimates of internal consistency for the scales.

Table 2 gives the rotated factor loadings of the scales on the 10 factors.⁶ The first factor appears to be a social desirability factor, with Scales 1-4 keyed for socially undesirable responses having high negative loadings and Scales 8, 9, and 12 keyed for socially desired responses having high positive loadings. It should be noted that there are scales at both

poles of the factor which are consistently keyed True. Furthermore, there are scales consistently keyed True which have relatively low loadings on the first factor. Scale 5, the Low Neutral scale, for example, has a loading of only -.17 and the two Couch and Keniston scales, Scales 15 and 16, have loadings of .14 and -.16, respectively, on the first factor, despite the fact that the items in all of these scales are also keyed True. It does not seem, therefore, that the first factor loadings can be accounted for in terms of a general tendency to acquiesce or answer True.

Scales 1-4 are keyed for socially undesirable responses and also for deviant True responses. All four scales have high negative loadings on the first factor. Scale 5, however, is also keyed for deviant True responses but not for socially desirable responses, since all of the items in this scale have neutral social desirability scale values, and this scale has a negligible loading of only -.17 on the first factor. Scale 10 is also keyed for deviant True responses and it also has a relatively low loading of only .25 on the first factor. Thus, it does not seem reasonable to interpret

⁶ The rotated factor loadings in Table 2 are *normalized* varimax loadings. The communalities for the 19 scales are: .89, .88, .80, .89, .94, .83, .83, .80, .82, .84, .87, .76, .92, .90, .95, .96, .98, .98, and .88, respectively. A 1-page table giving the unrotated factor loadings has been deposited with the American Documentation Institute. Order Document No. 7616 from ADI Auxiliary Publications Project, Photoduplication Service, Library of Congress, Washington, D. C. 20540. Remit in advance \$1.25 for microfilm or \$1.25 for photocopies and make checks payable to: Chief, Photoduplication Service, Library of Congress.

TABLE 2
ORTHOGONALLY ROTATED FACTOR LOADINGS OF THE 19 PERSONALITY SCALES

Scales	Factors									
	I	II	III	IV	V	VI	VII	VIII	IX	X
1. SUD ^a	-.98	.11	-.01	-.09	.05	-.03	.01	.00	.09	-.06
2. SUD	-.99	.04	.08	.02	.05	.03	-.02	.04	.01	-.01
3. SUD	-.91	-.04	-.11	.17	-.21	.13	-.03	-.07	-.01	.25
4. SUD	-.85	.05	-.02	.16	-.34	.19	.06	-.11	-.16	.23
5. Low Neutral	-.17	.28	.02	.03	.22	.32	-.86	.00	.01	-.01
6. Neutral	-.54	.52	-.15	.11	-.29	.36	-.34	-.20	-.13	.12
7. High Neutral	-.46	.36	.05	.33	-.21	.18	.11	-.04	-.64	.21
8. SD	.72	.51	-.12	.04	.02	.40	.06	-.18	-.04	.05
9. SD	.82	.42	.04	.19	-.09	.07	-.07	-.20	-.16	.18
10. SD Deviant	.25	.95	.07	-.03	.09	-.04	-.02	.10	-.08	-.09
11. SD Doubtful	.51	.82	-.12	.07	-.06	.00	.05	-.09	.16	.13
12. Edwards' SD	.86	.04	-.21	.03	.06	.09	.19	.04	.24	.32
13. M-C SD	.38	.18	.04	-.03	.91	-.02	-.01	-.01	.01	.00
14. Wiggins' Sd	.31	.35	-.01	-.09	.39	.78	.05	.04	-.05	.06
15. Couch-Keniston	.14	.27	-.13	.94	-.01	-.03	.00	.00	.00	.06
16. Couch-Keniston	-.16	.39	-.10	.45	.07	.41	-.06	-.11	.13	-.63
17. Total ?	-.12	.04	-.08	-.04	-.23	.13	.07	.95	.01	.00
18. P(T/?)	.18	.09	-.98	.01	-.01	.01	.00	.00	.00	.00
19. Consistency	.57	-.02	-.13	-.12	-.04	-.37	.20	-.08	-.64	-.19

^a See Footnote a, Table 1.

the first factor loadings in terms of a general tendency to give deviant True responses.

The scales with the highest loadings on Factor II are:

10. SD Deviant True	.95
11. SD Doubtful True	.82
6. Neutral True	.52
8. SD True	.51
9. SD True	.42

This factor appears to involve the tendency to give True responses to items with socially desirable scale values for which the modal response is False and also the tendency to give True response to items with socially desirable scale values which evoke doubts in many subjects as to whether or not the items accurately describe them. In addition, the Neutral scale containing items of high controversy keyed True has a positive loading on the factor. Scales 8 and 9 which contain items with socially desirable scale values keyed True also have positive loadings on this factor. At the same time, all of the scales containing items with socially undesirable scale values keyed True have very low load-

ings on the factor. A tentative interpretation of the factor is that it represents some combination of the tendency to give socially desirable responses and the tendency to acquiesce. In responding to items with socially desirable scale values the two tendencies should reinforce each other, since the acquiescent response is also the socially desirable response. On items with socially undesirable scale values the two tendencies should be in opposition, since the socially desirable response is False rather than True and this opposition may account for the low loadings of Scales 1-4 on the factor.

Only one scale, Scale 18, which is the conditional probability of a True response given that an item has been marked doubtful, has a high loading on Factor III. The fact that the mean probability of a keyed response on this scale is greater than .50 is consistent with results obtained with achievement tests in which it has been found that subjects are more likely to respond True than False when they are in doubt as to the correct response (Cronbach, 1946).

The Couch and Keniston HM and MM

scales have loadings of .94 and .45, respectively, on Factor IV. It is suggested that this factor may represent the tendency to agree with broad and general statements concerning beliefs and values about other people and the world in general. There is no evidence to indicate, however, that the kind of acquiescence represented by agreement with the items in the Couch and Keniston scales has any significant relationship to the tendency to agree with self-descriptive statements more directly related to what one does or how one behaves, i.e., the kinds of statements ordinarily included in personality trait scales. It is possible that the kind of acquiescence represented by the two Couch and Keniston scales is similar to the kind which is supposed to be present in the California F Scale (Bass, 1955; Cohn, 1953; Jackson & Messick, 1957; Messick & Jackson, 1957; Shelley, 1956).

The M-C *SD* scale has a loading of .91 on Factor V. This scale was modeled after the *L* scale of the MMPI and thus this factor may be tentatively interpreted as representing the tendency to give socially desirable but improbable responses.

Wiggins' *Sd* scale has a loading of .78 on Factor VI. An examination of the items contained in this scale shows that 20% are of the kind: "I would like to be a singer; I would like to be a nurse; I would like to be a soldier; I would like to be a private secretary; I very much like hunting;" and another 10% are of the kind: "I go to church almost every week; I pray several times a week; I read in the Bible several times a week." It is possible that Factor VI may represent in part the tendency to agree with statements of the kind noted.

The Low Neutral scale, Scale 5, has a loading of $-.86$ on Factor VII. All of the items are keyed for deviant True responses and all of the items have neutral social desirability scale values. Scale 5 has a low loading of $-.17$ on the first or social desirability factor. Factor VII may thus represent the tendency to give deviant True responses to personality items when the socially desirable response is not obvious.

Factor VIII is represented by Scale 17 which has a loading of .95 on the factor. No

other scale has a high loading on the factor and the factor may be interpreted tentatively as measuring the tendency to mark personality items as doubtful in self-description. This response set is also independent of social desirability tendencies.

Scale 7, in which neutral items are keyed for modal True responses, has a loading of $-.64$ on Factor IX and so also does Scale 19, the Consistency scale. It was suggested that both acquiescent and social desirability tendencies should tend to increase consistency of response. That consistency in response is associated with the tendency to give socially desirable responses is shown by the fact that Scale 19 has a positive loading on the first or social desirability factor. Factor IX seems to involve the tendency to give consistent responses and also the tendency to give modal True responses to neutral items.

The only scale with a substantial loading on Factor X is the Couch and Keniston MM scale which has a loading of $-.63$ on the factor. No attempt will be made to interpret this factor.

DISCUSSION

The results of this study show that it is possible to develop various experimental social desirability scales, along the lines suggested by Edwards (1957), using items from a source other than the MMPI. All of the experimental social desirability scales have a high degree of internal consistency and they also have high loadings on a common factor. Scales keyed for socially desirable responses have loadings on one pole of the factor and scales keyed for socially undesirable responses have loadings on the opposite pole of the factor. It is clear, however, that the *SD* scales developed by Crowne and Marlowe (1960) and by Wiggins (1959) are not measuring the same trait as that measured by the experimental social desirability scales and by Edwards' *SD* scale. The rationale underlying the M-C *SD* scale is that of the *L* scale of the MMPI, whereas Wiggins built his scale by finding MMPI items which differentiated between subjects given the MMPI under standard instructions and under instructions to fake good. There is, of course, no reason why social desirability scales developed by dif-

ferent techniques should measure the same trait.

Three types of scales keyed for deviant True responses were investigated. In one type, the keying was for deviant True responses to items with socially undesirable scale values; in another the keying was for deviant True responses to neutral items; and in the third the keying was for deviant True responses to items with socially desirable scale values. These three types of deviant True scales did not have their highest loadings on a single common factor but instead each type had its highest loading on a different factor. This result suggests that there is no general tendency to give deviant True responses to items which operate independently of the social desirability scale values of the items.

The tendency to give deviant True responses to MMPI items has been found to be uncorrelated with the tendency to give deviant False responses (Barnes, 1956b; Wiggins, 1962). In the present study, however, it was shown that these two tendencies are not uncorrelated if the items involved are those with socially undesirable scale values keyed for deviant True responses and those with socially desirable scale values keyed for deviant False responses. On the other hand, the tendency to give deviant True responses to neutral items, as measured by Scale 5, is relatively uncorrelated with the tendency to give deviant False responses to neutral items as measured by Scale 7 (reflected), the correlation between these two scales being only $-.09$.

The results of this study have implications for the development of scales designed to measure acquiescent tendencies. Edwards (1957) has suggested that responses to items with neutral social desirability scale values should be relatively uninfluenced by social desirability tendencies. If, in addition, the items are of high controversiality, with True and False responses almost equally probable, then one might expect that such a pool of items would be ideal for measuring acquiescent tendencies. The present study shows, however, that Scale 6, which possesses these ideal properties, has a loading of $-.54$ on the first or social desirability factor. Thus, merely selecting neutral items or items of high con-

troversiality or items with both properties is not sufficient to obtain a measure of acquiescence free from social desirability tendencies.

In the present study, the only neutral scale which has a low loading on the first or social desirability factor is Scale 5 in which the items are keyed for deviant True responses. Scale 5 has relatively low correlations with all of the social desirability and social undesirability scales, the highest correlation being $.22$ with Scale 4. Scale 5 is a relatively pure measure of Factor VII. If this scale is measuring acquiescent tendencies, then it must be admitted that this response set is rather specific and is unrelated to performance on the other scales investigated.

It has been suggested by Cronbach (1946) that acquiescent tendencies are most apt to be elicited by achievement items when a subject does not know or is in doubt as to the correct response to the item. A similar measure for personality items was included in the present study. This measure is Scale 18 which yields for each subject a conditional probability of a True response for those items he marked doubtful. This response set is also rather specific and independent of social desirability tendencies. Scores on Scale 18 are, however, also uncorrelated with scores on the three neutral scales included in the present study. If Scale 18 is regarded as measuring the tendency to acquiesce, then it is a different form of acquiescence than that being measured by Scale 5 or the other two neutral scales.

Still another scale which has been regarded as measuring acquiescent tendencies is the Couch and Keniston HM scale, Scale 15. This scale also has a low loading on the first or social desirability factor and has its highest loading on Factor IV, a factor on which none of the other proposed measures of acquiescence has a high loading.

Of the six proposed measures of acquiescent tendencies, Scales 5-7, Scales 15 and 16, and Scale 18, each has its highest factor loading on a different factor. In view of these results, it would appear that considerable attention needs to be given to the problem of the meaning of acquiescence and how to measure it, before one can conclude that it is a general response set influencing scores

on all personality scales to the degree to which there is an imbalance in the True-False keying of the scales.

REFERENCES

- BARNES, E. H. Factors, response bias, and the MMPI. *J. consult. Psychol.*, 1956, 20, 419-421. (a)
- BARNES, E. H. Response bias and the MMPI. *J. consult. Psychol.*, 1956, 20, 371-374. (b)
- BASS, B. M. Authoritarianism or acquiescence? *J. abnorm. soc. Psychol.*, 1955, 51, 616-623.
- BURKET, G. An analytic iterative rotation program for the IBM Type 709. Seattle: University of Washington, 1961. (Mimeo)
- COHN, T. S. The relationship of the F scale to a response set to answer positively. *Amer. Psychologist*, 1953, 8, 335.
- COUCH, A., & KENISTON, K. Yeasayers and nay-sayers: Agreeing response set as a personality variable. *J. abnorm. soc. Psychol.*, 1960, 60, 151-174.
- CRONBACH, L. J. Response sets and test validity. *Educ. psychol. Measmt.*, 1946, 6, 475-494.
- CRONBACH, L. J. Further evidence on response sets and test design. *Educ. psychol. Measmt.*, 1950, 10, 3-31.
- CROWNE, D. P., & MARLOWE, D. A new scale of social desirability independent of psychopathology. *J. consult. Psychol.*, 1960, 24, 349-354.
- EDWARDS, A. L. The relationship between the judged desirability of a trait and the probability that the trait will be endorsed. *J. appl. Psychol.*, 1953, 37, 90-93.
- EDWARDS, A. L. *The social desirability variable in personality assessment and research*. New York: Dryden Press, 1957.
- EDWARDS, A. L. Social desirability or acquiescence in the MMPI? A case study with the SD scale. *J. abnorm. soc. Psychol.*, 1961, 63, 351-359.
- EDWARDS, A. L., & DIERS, CAROL J. Social desirability and the factorial interpretation of the MMPI. *Educ. psychol. Measmt.*, 1962, 22, 501-509.
- EDWARDS, A. L., DIERS, CAROL J., & WALKER, J. N. Response sets and factor loadings on sixty-one personality scales. *J. appl. Psychol.*, 1962, 46, 220-225.
- EDWARDS, A. L., & HEATHERS, LOUISE B. The first factor of the MMPI: Social desirability or ego strength? *J. consult. Psychol.*, 1962, 26, 99-100.
- EDWARDS, A. L., HEATHERS, LOUISE B., & FORDYCE, W. E. Correlations of new MMPI scales with Edwards' SD scale. *J. clin. Psychol.*, 1960, 16, 26-29.
- EDWARDS, A. L., & WALSH, J. A. The relationship between the intensity of the social desirability keying of a scale and the correlation of the scale with Edwards' SD scale and the first factor loadings of the scale. *J. clin. Psychol.*, 1963, 19, 200-203.
- FRICKE, B. G. Response set as a suppressor variable in the OASIS and the MMPI. *J. consult. Psychol.*, 1956, 20, 161-169.
- HANLEY, C. Social desirability and responses to items from three MMPI scales: D, Sc, and K. *J. appl. Psychol.*, 1956, 40, 324-328.
- HATHAWAY, S. R., & MCKINLEY, J. C. *Manual for the Minnesota Multiphasic Personality Inventory*. New York: Psychological Corporation, 1951.
- JACKSON, D. N., & MESSICK, S. A note on ethnocentrism and acquiescent response sets. *J. abnorm. soc. Psychol.*, 1957, 54, 132-134.
- JACKSON, D. N., & MESSICK, S. Content and style in personality assessment. *Psychol. Bull.*, 1958, 55, 243-252.
- JACKSON, D. N., & MESSICK, S. Acquiescence and desirability as response determinants on the MMPI. *Educ. psychol. Measmt.*, 1961, 21, 771-790.
- McGEE, R. K. Response style as a personality variable: By what criterion? *Psychol. Bull.*, 1962, 59, 284-295.
- MESSICK, S., & JACKSON, D. N. Authoritarianism and acquiescence in Bass's data. *J. abnorm. soc. Psychol.*, 1957, 54, 424-426.
- MESSICK, S., & JACKSON, D. N. Acquiescence and the factorial interpretation of the MMPI. *Psychol. Bull.*, 1961, 58, 299-304.
- NUSSBAUM, F., MARSHALL, D. A., & ROOSEN-RUNGE, P. Symmetric correlation and principal axis factor analysis package program for the IBM Type 709. Seattle: University of Washington, 1961. (Mimeo)
- SHELLEY, H. P. Response set and the California attitude scales. *Educ. psychol. Measmt.*, 1956, 16, 63-67.
- WELSH, G. A. An anxiety index and an internalization ratio for the MMPI. *J. consult. Psychol.*, 1952, 16, 65-72.
- WIGGINS, J. S. Interrelationships among MMPI measures of dissimulation under standard and social desirability instructions. *J. consult. Psychol.*, 1959, 23, 419-427.
- WIGGINS, J. S. Strategic, method, and stylistic variance in the MMPI. *Psychol. Bull.*, 1962, 59, 224-242.

(Received September 27, 1962)

A MODIFIED MODEL FOR TEST VALIDATION AND SELECTION RESEARCH¹

MARVIN D. DUNNETTE

University of Minnesota

It is argued that the classic prediction model is grossly oversimplified and has resulted in corresponding oversimplifications in the design of most validation studies. A modified and more complex prediction model is presented. Implications for future validation research are discussed in the context of the kinds of behaviors to be predicted, the necessity for investigating heteroscedastic and nonlinear relationships, and the important advantages in prediction which may be realized by discovering homogeneous subsets of jobs, tests, people, and behaviors within which prediction equations may be developed and cross-validated.

Nearly 35 years ago, Clark Hull (1928) discussed the level of forecasting efficiency shown by the so-called modern tests of the time. He noted that the upper limit for tests was represented by validity coefficients of about .50 corresponding to a forecasting efficiency of only 13%. He regarded the region of forecasting efficiency lying above this point as being inaccessible to the test batteries of the day, and he viewed with pessimism the use of test batteries for predicting occupational criteria. Hull, of course, failed to emphasize that the accuracy of practical decisions might better be assessed against zones of behavior (e.g., passing versus failing in a training program) rather than against the metrical continuum assumed in the calculation of his index of forecasting efficiency. Further, he gave no attention to the varying effects of different selection ratios on the accuracies obtainable with even rather low correlation coefficients. Even so, we should be somewhat dismayed by the fact that today our tests have still not penetrated the region of inaccessibility defined so long ago by Hull. Ghiselli's (1955) comprehensive review of both published and unpublished studies showed average validities ranging in the .30s and low .40s; an average validity of .50 or above was a distinct rarity. These low validities have apparently led many psychologists to become disenchanted with test and selection research. Some have disappeared into

other endeavors such as the study of group influences, interaction patterns, and the like. Others have sought refuge in the hypothesis testing models of statistical inference and have implied validity for tests showing *statistically* (but often not *practically*) significant differences between contrasting groups (see Dunnette & Kirchner, 1962). Nunnally (1960) comments:

We should not feel proud when we see the psychologist smile and say "the correlation is significant beyond the .01 level." Perhaps that is the most he can say, but he has no reason to smile [p. 649].

Even less defensible, perhaps, has been the tendency for many to persist in doing selection *without* conducting selection research or test validation. The ordinary defenses for such practice run the gamut—from claiming near miracles of clinical insight in personnel assessment to the recounting of anecdotes about instances of selective accuracy (counting the "hits" and forgetting the "misses") and finally to the old cliché that "management is well-satisfied with the methods being employed." We cannot and should not try to avoid the fact that the statistics of selection (i.e., validity coefficients) are far from gratifying and offer little support to anyone claiming to do *much* better than chance in the selection process.

It seems wise, therefore, to discuss the possibility of improving our batting average in test validation and selection research. Selection programs will go on—with or without psychologists—but I believe we now have

¹ This paper was read at the seventieth annual convention of the American Psychological Association held in St. Louis in the fall of 1962.

the capability for penetrating the region of inaccessibility outlined by Hull.

First, let us examine the classic validation or prediction model. This model has sought simply to link predictors, on the one hand, with criteria, on the other, through a simple index of relationship, the correlation coefficient. Such a simple linkage of predictors and criteria is grossly oversimplified in comparison with the complexities actually involved in predicting human behavior. Most competent investigators readily recognize this fact and design their validation studies to take account of the possible complexities—job differences, criterion differences, etc.—present in the prediction situation. Even so, the appealing simplicity, false though it is, of the classic model has led many researchers to be satisfied with a correspondingly simplified design for conducting selection research. Thus, the usual validation effort has ignored the events—on the job behavior, situational differences, dynamic factors influencing definitions of success, etc.—intervening between predictor and criterion behavior. I believe that the lure of this seemingly simple model is, to a great extent, responsible for the low order of validities reported in the Ghiselli (1955) review. It is noteworthy that the studies reviewed by Ghiselli show no typical level of prediction for any given test or type of job. In fact, there seems to be little consistency among various studies using similar tests and purporting to predict similar criteria. The review also suggests that the magnitude of validity coefficients is inversely proportional to the sample size employed in the studies. This can perhaps be explained, in part, by sampling error, but it may also be due to the relatively greater homogeneity possible within smaller groups of subjects. It appears, in other words, that the varying levels of prediction shown by the various studies are related somehow to the appropriateness (or lack thereof) of the classic prediction model for the particular set of conditions in the study being reported. It seems wise, therefore, to consider a prediction model which more fully presents the complexities which are only implied by the classic model.

Guetzkow and Forehand (1961) have suggested a modification of the classic validation model which provides a richer schematization for prediction research and which offers important implications for the direction of future research. Their model along with certain additional modifications is shown in Figure 1. Note that the modified prediction model takes account of the complex interactions which may occur between predictors and various predictor combinations, different groups (or types) of individuals, different behaviors on the job, and the consequences of these behaviors relative to the goals of the organization. The model permits the possibility of predictors being differentially useful for predicting the behaviors of different subsets of individuals. Further, it shows that similar job behaviors may be predictable by quite different patterns of interaction between groupings of predictors and individuals or even that the same level of performance on predictors can lead to substantially different patterns of job behavior for different individuals. Finally, the model recognizes the annoying reality that the same or similar job behaviors can, after passing through the situational filter, lead to quite different organizational consequences.

This modified and more complex prediction model leads to a number of important considerations involving the emphases to be followed by future validation research:

First, we must be willing to back off a step or two from global measures of occupational effectiveness—ratings, volume of output, and other so-called criteria of organizational worth, and do a more careful job of studying actual job behavior—with particular focus on behavioral or stylistic variations among different individuals with the same jobs. Most previous validation research has been overly concerned with predicting organizational consequences without first determining the nature of possible linkages between such consequences and differences in actual job behavior. It is true that industrial psychologists should continue to be concerned about predicting organizational consequences. Certainly, the modified model implies no lessening of such an interest. What

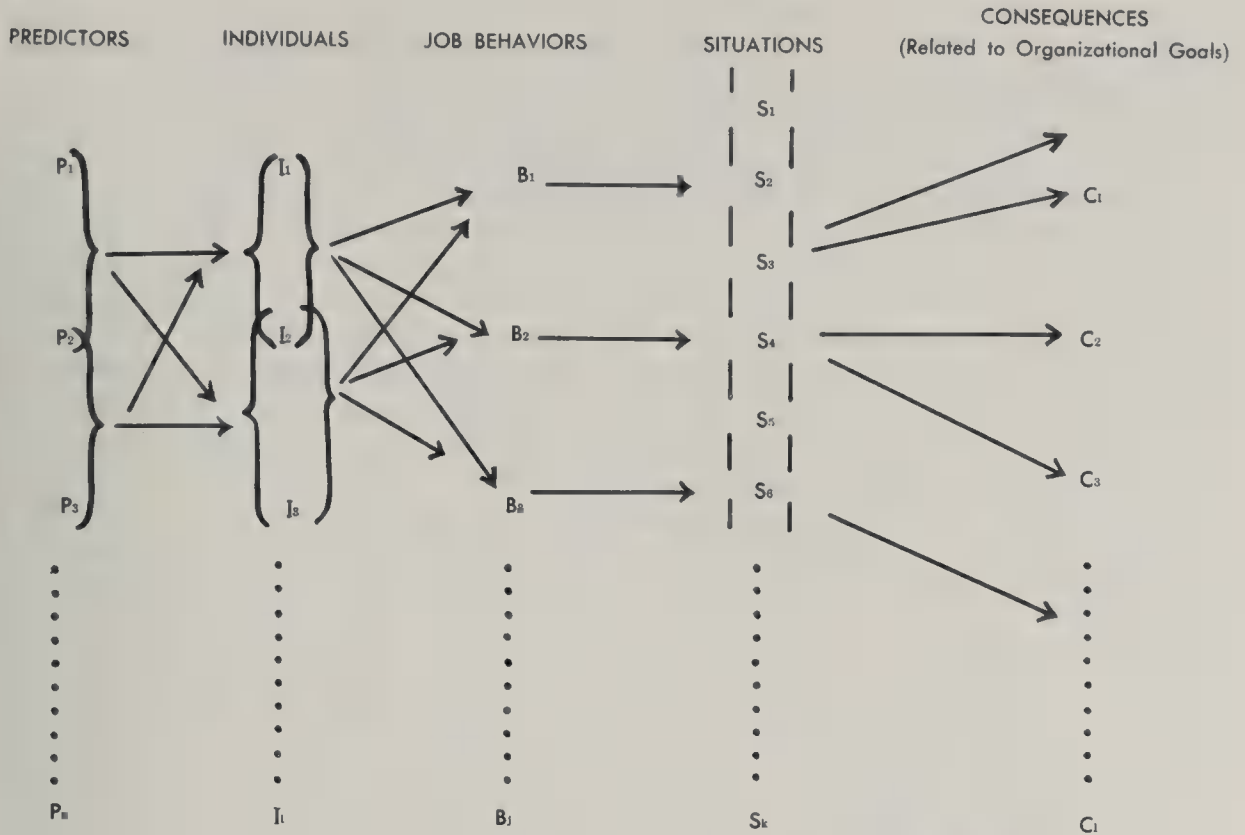


FIG. 1. A modified model for test validation and selection research.

is hoped, however, is that the more careful analysis of the behavioral correlates of differences in organizational consequences will lead to broader understanding of them and, eventually, to their more accurate prediction.

Secondly, as implied by the point just made, the modified model demands that we give up our worship of *the* criterion (Dunnette, 1963). I believe that our concept of *the* criterion has suggested the existence of some single, all encompassing measure of occupational success against which predictors must be compared. Our modified model demands that we work with multiple measures of individual behavior and organizational consequences. I suggest therefore that we cease talking about *the* criterion problem and that we discard the notion of a so-called ultimate criterion. Such action should result in a research emphasis which will be less restrictive and less simple-minded and more aware of the necessity of analyzing and predicting the many facets of occupational success.

Thirdly, the modified model implies nothing

concerning the form of the relationships to be expected. One of the unfortunate consequences of utilizing the classic validation model was its overemphasis on the correlation coefficient as almost the sole statistic of validation research. The notion of a simple linkage between predictor and criterion led easily to the equally simple assumption of the applicability of the linear, homoscedastic model for expressing the magnitude of relationships. Kahneman and Ghiselli (1962), in investigating relationships between 60 aptitude variables and various criteria, showed that 40% of the scatter diagrams departed significantly from the linear, homoscedastic model, and 90% of these departures held up on cross-validation. This is an important finding for it points up the necessity in future validation research of adopting a methodology taking account of the very great likelihood of nonlinear, heteroscedastic models. Our more complex prediction model, focusing as it does on the complex linkages between predictors and consequences, implies also the necessity of adopting more complex

and sophisticated tools of analysis in studying these linkages.

Fourth, and most obviously, our modified model demands that we develop a sort of typology for classifying people, tests, job situations, and behaviors according to their relative predictability. Future validation research must define the unique conditions under which certain predictors may be used for certain jobs and for certain purposes. Research studies should, therefore, be devoted to the definition of homogeneous subsets within which appropriate prediction equations may be developed and cross-validated. This idea is not particularly startling nor even new. But it has *not* been applied widely in the conduct of selection research. The modified model rather explicitly directs us to carry out such subgrouping studies in order to learn more about the complex linkages between predictors and consequences. Fortunately several studies already are available which confirm the advantages of studying differential patterns of validity for various subgroups. A brief review of some of these research approaches should illustrate the utility of applying our more complicated model to validation research.

With respect to job groupings, Dunnette and Kirchner (Dunnette, 1958; Dunnette & Kirchner, 1958, 1960) have studied the different patterns of validities obtained when careful techniques of job analysis are used to discover groupings of jobs which are relatively homogeneous in terms of actual responsibilities. Substantially different validities were obtained for engineers grouped according to functional similarities (research, development, production, and sales), salesmen (industrial and retail), and clerical employees (stenographers and clerk typists). These studies highlight the necessity of studying job differences and the differential predictability of effectiveness in various job groupings. More generally, an emphasis on the varying predictability of different job activities is inherent in the methods of synthetic validity (Balma, Ghiselli, McCormick, Primoff, & Griffin, 1959) and in the use of the *J* coefficient developed by Primoff (1955).

Everyone recognizes the possibility of situ-

ational effects on the validity of psychological predictions, but there is a paucity of research designed to estimate systematically the magnitude of such effects. Perhaps the best example of such research is provided by Vroom (1960). He showed that various aptitude tests (verbal and nonverbal reasoning, arithmetic reasoning) predicted ratings of job success most effectively for persons who were highly motivated. Job effectiveness in nonmotivating situations showed either no relationship or negative relationships with tested abilities. In a second study with Mann (Vroom & Mann, 1960), it was shown that the size of work groups strongly influenced employee attitudes toward their supervisors. Employees in small groups preferred democratic or equalitarian supervisors; employees in large work groups preferred authoritarian supervisors. In a significant series of studies, Porter (1962) is also investigating situational factors such as hierarchical level, firm size, and job function as they affect managerial perceptions of their jobs. More emphasis needs to be given to these and other situational factors in validation studies, particularly as they serve to operate as moderating variables (Saunders, 1956) in behavioral predictions.

Many studies have shown different validities for different subgroups of individuals. For example, Seashore (1961) summarized a vast number of scholastic success studies which show almost uniformly that the grades of women (in both high school and college) are significantly more predictable than those of men. It is also well established that differing patterns of validity are typically obtained for subgroups differing in amounts of education and/or years of job experience. It may seem obvious that such factors as sex, education, and experience provide useful moderating variables in validation research. However, researchers also have identified variables which are much less *obvious* but which *do* make substantial differences in the patterns and magnitudes of validities obtained. For example, Grooms and Endler (1960) showed that the grades of anxious college students were much more predictable ($r = .63$) with aptitude and achievement

measures than were the grades of nonanxious students ($r = .19$); and Frederiksen, Melville, and Gilbert (Frederiksen & Gilbert, 1960; Frederiksen & Melville, 1954) have shown that interest in engineering (as measured by the Strong test) has a higher validity for predicting grades for noncompulsive engineers than for compulsive ones. Berdie (1961) showed that the grades of engineering students with relatively consistent scores on an algebra test were more predictable from the total test score than were the grades of students with less consistent scores.² Ghiselli (1956, 1962) has developed a method for dividing persons, on the basis of a screening test, into more and less predictable subgroups. The advantage of his method is that no a priori basis is necessary for the identification of subgroups; the method depends simply on the development of one or more predictor tests to facilitate the subgrouping process.

The identification of more and less predictable subgroups of persons, whether based on logical factors (such as sex, education, or experience) or on methods such as those employed by Berdie and Ghiselli, places a special burden on the investigator to demonstrate the stability of his results. Although the studies cited above were cross-validated (i.e., checked on hold-out groups), the validity generalization and/or extension of such results has not often been measured. This needs to be done. The results so far reported with these methods are promising indeed, but they will take on greatly added significance when it is demonstrated that they hold up over time.

Less research has been directed at identifying subsets of predictors showing differential patterns of validity. However, Ghiselli (1960, 1962) has also contributed methodology in this area and has succeeded in significantly enhancing prediction by identifying, again through the development and use of screening tests, the particular predictor

which will do the most valid job for each individual.

General approaches to the development of "types" have been made by a number of investigators. Gaier and Lee (1953) and Cronbach and Gleser (1953) summarize a variety of methods of assessing profile similarity and conclude that available indexes are simply variants of the general Pythagorean formula for the linear distance between two points in n -dimensional space. Lykken (1956) has questioned the psychological meaning of such "geometric similarity" and he proposes a method of actuarial pattern analysis which requires no assumptions concerning the form of the distribution and which defines similarity in psychological rather than geometric terms. His method consists simply of investigating criterial outcomes for subjects classified together into cells on the basis of similar test scores. In a recent study, he and Rose (Lykken & Rose, in press) demonstrate that the method is more accurate in discriminating between neurotics and psychotics on the basis of MMPI scores than either clinicians' judgments or a statistical technique based on equations derived from a discriminant function analysis. Lykken's method of actuarial pattern analysis is the same as Toops' (1959) method of developing subgroups or "ulstriths" based on biographical and test similarities and then writing different prediction equations for each of the subgroups so identified. It is interesting to note that computers have now given us the capability for carrying out many of Toops' suggestions—which at one time were regarded as wild-eyed, idealistic, and unrealistic. McQuitty (1957, 1960, 1961) also has developed methods for discovering the diagnostic and predictive significance of various response patterns. His techniques, in addition to the methods proposed by Lykken and Toops, constitute the most extensive attack made to date on the problem of developing differentially predictable subsets or types.

These studies and methods mark the bare beginnings of efforts to take account of complexities which have been ignored by the oversimplified prediction model of the past. It appears that subgrouping of tests, people,

² The algebra test of 100 items was divided into 10 subtests of equal difficulty. The measure of consistency for each student was simply the sum of squares of the deviations of his 10 scores from his mean score on all 10 subtests.

jobs, situations, and consequences is necessary to a thorough understanding of what is going on in a prediction situation. The widespread acceptance of the modified model which we have been discussing should lead to a new and refreshing series of questions about problems of selection and placement. Instead of asking whether or not a particular selection technique (test, interview, or what have you) is any good, we will ask under *what circumstances* different techniques may be useful. What sorts of persons should be screened with each of the methods available, and how may the various subgroups of persons be identified and assigned to optimal screening devices? Finally, what job behaviors may be expected of various people and how may these behaviors be expected to aid or to detract from accomplishing different organizational objectives which may, in turn, vary according to different value systems and preferred outcomes?

What are the implications of these trends for the selection function in industry? Primarily, I believe they suggest the possibility of a new kind of selection process in the firm of the future. The selection expert of tomorrow will no longer be attempting to utilize the same procedure for all his selection problems. Instead, he will be armed with an array of prediction equations. He will have developed, through research, a wealth of evidence showing the patterns of validities for different linkages in the modified prediction model—for different predictors, candidates, jobs, and criteria. He will be a flexible operator, attentive always to the accumulating information on any given candidate, and ready to apply, at each stage, the tests and procedures shown to be optimal.

REFERENCES

- BALMA, M. J., GHISELLI, E. E., MCCORMICK, E. J., PRIMOFF, E. S., & GRIFFIN, C. H. The development of processes for indirect or synthetic validity: A symposium. *Personnel Psychol.*, 1959, 12, 395-400.
- BIRDIE, R. F. Intra-individual variability and predictability. *Educ. psychol. Measmt.*, 1961, 21, 663-676.
- CRONBACH, L. J., & GLESER, GOLDINE. Assessing similarity between profiles. *Psychol. Bull.*, 1953, 50, 456-473.
- DUNNETTE, M. D. Validity of interviewer's ratings and psychological tests for predicting the job effectiveness of engineers. St. Paul: Minnesota Mining and Manufacturing Company, 1958. (Mimeo)
- DUNNETTE, M. D. A note on the criterion. *J. appl. Psychol.*, 1963, 47, 251-254.
- DUNNETTE, M. D., & KIRCHNER, W. K. Validation of psychological tests in industry. *Personnel Admin.*, 1958, 21, 20-27.
- DUNNETTE, M. D., & KIRCHNER, W. K. Psychological test differences between industrial salesmen and retail salesmen. *J. appl. Psychol.*, 1960, 44, 121-125.
- DUNNETTE, M. D., & KIRCHNER, W. K. Validities, vectors, and verities. *J. appl. Psychol.*, 1962, 46, 296-299.
- FREDERIKSEN, N., & GILBERT, A. C. Replication of a study of differential predictability. *Educ. psychol. Measmt.*, 1960, 20, 759-767.
- FREDERIKSEN, N., & MELVILLE, S. D. Differential predictability in the use of test scores. *Educ. psychol. Measmt.*, 1954, 14, 647-656.
- GAIER, E. L., & LEE, MARILYN. Pattern analysis: The configural approach to predictive measurement. *Psychol. Bull.*, 1953, 50, 140-148.
- GHISELLI, E. E. The measurement of occupational aptitude. Berkeley: Univer. California Press, 1955.
- GHISELLI, E. E. Differentiation of individuals in terms of their predictability. *J. appl. Psychol.*, 1956, 40, 374-377.
- GHISELLI, E. E. Differentiation of tests in terms of the accuracy with which they predict for a given individual. *Educ. psychol. Measmt.*, 1960, 20, 675-684.
- GHISELLI, E. E. The prediction of predictability and the predictability of prediction. Paper read at American Psychological Association, St. Louis, September 1962.
- GROOMS, R. R., & ENDLER, N. S. The effect of anxiety on academic achievement. *J. educ. Psychol.*, 1960, 51, 299-304.
- GUETZKOW, H., & FOREHAND, G. A. A research strategy for partial knowledge useful in the selection of executives. In R. Taguiri (Ed.), *Research needs in executive selection*. Boston: Harvard Graduate School of Business Administration, 1961.
- HULL, C. L. Aptitude testing. Yonkers, N. Y.: World Book, 1928.
- KAHNEMAN, D., & GHISELLI, E. E. Validity and non-linear heteroscedastic models. *Personnel Psychol.*, 1962, 15, 1-11.
- LYKKEN, D. T. A method of actuarial pattern analysis. *Psychol. Bull.*, 1956, 53, 102-107.
- LYKKEN, D. T., & ROSE, R. J. Psychological prediction from actuarial tables. *J. clin. Psychol.*, in press.
- MCQUITTY, L. L. Isolating predictor patterns associated with major criterion patterns. *Educ. psychol. Measmt.*, 1957, 17, 3-42.

- McQUITTY, L. L. Hierarchical linkage analysis for the isolation of types. *Educ. psychol. Measmt.*, 1960, 20, 55-67.
- McQUITTY, L. L. A method for selecting patterns to differentiate categories of people. *Educ. psychol. Measmt.*, 1961, 21, 85-94.
- NUNNALLY, J. The place of statistics in psychology. *Educ. psychol. Measmt.*, 1960, 20, 641-650.
- PORTER, L. W. Some recent explorations in the study of management attitudes. Paper read at American Psychological Association, St. Louis, September 1962.
- PRIMOFF, E. S. *Test selection by job analysis*. Washington, D. C.: United States Civil Service Commission, Test Development Section, 1955.
- SAUNDERS, D. R. Moderator variables in prediction. *Educ. psychol. Measmt.*, 1956, 16, 209-222.
- SEASHORE, H. G. Women are more predictable than men. Presidential address, Division 17, American Psychological Association, New York, September 1961.
- TOOPS, H. A. A research utopia in industrial psychology. *Personnel Psychol.*, 1959, 12, 189-227.
- VROOM, V. H. *Some personality determinants of the effects of participation*. Englewood Cliffs, N. J.: Prentice-Hall, 1960.
- VROOM, V. H., & MANN, F. C. Leader authoritarianism and employee attitudes. *Personnel Psychol.*, 1960, 13, 125-139.

(Received October 1, 1962)

EMOTIONAL AROUSAL AND TASK PERFORMANCE¹

BIBB LATANÉ

AND

A. JOHN ARROWOOD

Columbia University

University of Toronto

42 male Ss were exposed to either a hostility inducing or a neutral confederate and then tested on both stereotyped and nonstereotyped forms of a simple lever pressing task. As predicted, emotional arousal (hostility) had no effect on performance of the relatively stereotyped task but led to a considerable though transitory drop in performance after changeover to a nonstereotyped task requiring concentration. This result corroborates in a laboratory setting previous reports of research done in the field.

Production during industrial changeover is characteristically unpredictable. Although on occasion new work procedures can be introduced with little difficulty, more often productivity will drop precipitously and approach desired levels only days or even weeks after changeover. In an attempt to examine some of the psychological factors which may be related to this effect, Schachter, Willerman, Festinger, and Hyman (1961) conducted field studies designed to assess the relationship between emotional disruption and productivity during and between industrial changeovers. They suggest that the typical industrial assembly operation can be characterized as a stereotyped or automatized form of behavior requiring neither concentration, attention, nor thought. Assembly operations during and immediately following changeover, however, cannot be considered stereotyped; even the smallest change in work procedure will require attention and concentration until the new task is mastered. Schachter et al. propose that the effect of emotional disruption upon productivity is related to the extent to which a given assembly operation is stereotyped. Emotional disruption, they contend, will have little or no effect on a worker's performance at a well-learned and practiced task, but will interfere markedly with performance during task changeover.

To test this proposition, pairs of experienced production lines in an appliance factory were matched according to job characteristics. Workers in one of each pair of lines were subjected to a stream of harassment, abuse, and irritation extending over a period of several weeks and presumably creating emotional arousal of a negative sort. Workers in the corresponding "control" lines were shielded as much as possible from annoyance. Records kept during this phase of the study evidence no signs of any effects of the emotional manipulation on stereotyped performance. During the second phase of the study, all lines underwent major job changeovers of the sort common to most industry. Those lines which had been subjected to the emotional manipulation showed a larger and longer lasting disruption of production than did their matched control lines.

Although the results of this study are positive and intuitively convincing, they are subject to the same limitations which befall field studies in general. Perhaps the most critical of these is that, even with the expenditure of a great deal of time and money (and lost production), it was possible to collect only a few cases. In such a situation, the use of statistical techniques to evaluate the evidence is precluded and the possibility that chance is the major factor responsible for the results is increased.

The present experiment was designed to provide a further test of the Schachter et al. proposition in a laboratory situation. The design of the experiment was similar to that employed in the field study described above. Subjects were given intensive training on one

¹ This research was carried out at the Laboratory for Research in Social Relations at the University of Minnesota. We wish to express our great appreciation to Barry Schuler and Barbara Koltes who served as experimenter and instigator, respectively, and who aided in the design of the experiment and the analysis of the data. The General Electric Foundation provided financial support for this study.

task, subjected to emotional arousal, and then transferred to a somewhat different task. It was predicted that emotional arousal would have no effect on the performance of the first task, but that when the task was changed so that stereotyped behavior was no longer possible, emotionally aroused subjects would suffer a greater decrement in performance than would nonaroused subjects.

METHOD

Subjects

Forty-two men enrolled in introductory psychology courses at the University of Minnesota volunteered to serve in the experiment. They received extra credit toward their course grade in return for their participation.

Task

The task for all subjects throughout the experiment consisted of pressing switches in a simple repetitive sequence. Subjects were seated in front of a console on which was mounted an array of three pilot lights and three switches. The pilot lights were lit sequentially and served to define the correct response. All subjects were carefully instructed in the light sequence since their task, both before and after changeover, was to anticipate the oncoming light by pressing the appropriate switch. In the first part of the experiment, subjects were instructed to press the switch in the same *position* (right, middle, or left) as the light which was due to appear. After task changeover, their job was to press the switch of the same *color* (red, green, or blue) as the light which was to come on. Subjects could work at their own pace since a new light would come on whenever a switch was pressed, whether correctly or incorrectly.

Procedure

Premanipulation Period. Subjects were given 12 minutes of training on the first task, which consisted of anticipating which light would come on next by pressing the switch underneath that light. A minute-by-minute record was kept of the total number of times each subject pressed any switch and of the total number of correct presses.

Emotional Instigation. At the beginning of the experimental hour, each subject filled out a short questionnaire introduced by the experimenter as part of another study being conducted by a graduate student and of no direct relevance to the task at hand. After 12 minutes of training on Task I, an attractive young woman, ostensibly this graduate student, entered the room and asked for the opportunity to clarify some of the subject's questionnaire responses. The experimenter reluctantly agreed, but instructed the subject not to stop work during this interruption. In the Hostility condition, after look-

ing through the questionnaire and asking a few personal questions, the instigator launched into a personal attack on the hapless subject's age, grades, questionnaire responses, and general qualifications, and then stalked out of the room with a parting thrust at the subject's taste in clothes. The subjects in the Control condition filled out the same questionnaire and were exposed to the same instigator, but the instigator's behavior was considerably different. Although the same personal questions were asked, the instigator regarded the answers with mild approval. Each manipulation took approximately 2 minutes.²

Continuation of Task I. After the instigator had left the room, the subjects continued working on Task I for an additional 3 minutes. Productivity during the interval was used to index the effect of emotional arousal on the performance of a relatively stereotyped task.

Task Changeover. Seventeen minutes after starting on Task I, all the subjects were given a new task on which to work. Whereas in Task I they had been required to press the switch immediately *below* the light which was to come on, their job now was to press the switch of the same *color* as the light which was due to appear. Since the sequence of lights remained the same, the new task required only a slight change in the sequence of hand movements.

RESULTS

Emotional Manipulation

At the end of the experimental session, the subjects filled out self-report scales on the extent to which they felt anxious, tense, irritated, content, and happy. Four of these measures failed to discriminate significantly between the Control and Hostility groups. The groups, however, differed significantly ($p = .05$ by t test)³ on the tenseness scale. Since the Hostility group was the more tense at the end of the experimental session, there is at least minimal evidence that the experimental manipulation was effective. Spontaneous comments made by the subjects in the Hostility condition during or immediately after the emotional manipulation are perhaps

² An attempt was made to create a third experimental group in which subjects were subjected to a "euphoria" manipulation. Since self-report measures, observation, and postexperimental discussion with the subjects indicated that the manipulation was not successful in arousing happiness, and since these subjects did not differ significantly from the Control group in performance, this condition was dropped from the experiment.

³ All p values reported in this paper are two-tailed.

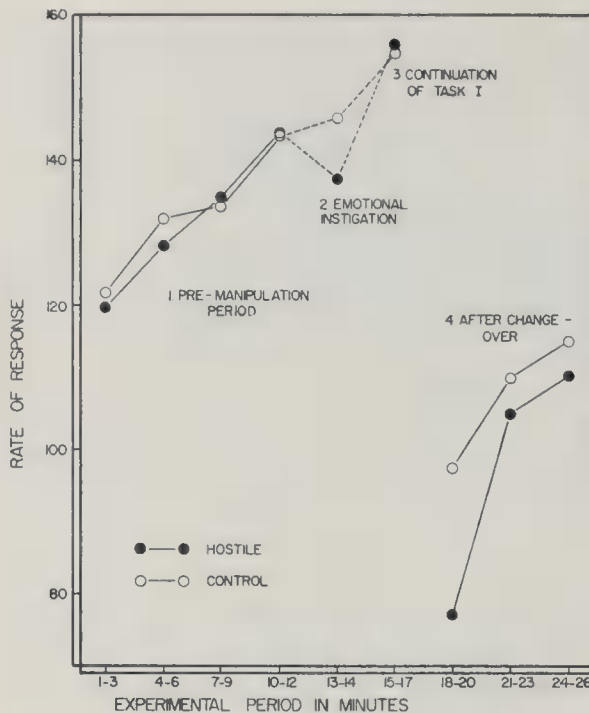


FIG. 1. Total responses by minute of experimental period.

more revealing than the self-report questionnaire completed at the end of the experiment. One subject was so ungallant as to tell the young lady "get lost, you're bothering me" while other subjects were moved to wonder "what's eating that old bitch?" or "what the hell did she expect, a tuxedo?" No such reactions were forthcoming from subjects in the Control condition.

Productivity

Figure 1 presents the mean total number of responses made by each of the experimental groups during the experimental session. The results will be discussed separately for each of four periods.

Premanipulation Period. The subjects worked for 12 minutes on Task I before the emotional manipulation was introduced. It can be seen from Figure 1 that the experimental groups did not differ significantly during this period.

Emotional Instigation. During this 2-minute period, the subjects were working under the distraction of an attempt to manipulate their emotional state. It can be seen in Figure 1 that the subjects in the Hostility group performed at a somewhat slower rate than did

their Control counterparts. Although this difference is not statistically significant, it is considerably larger than any of the preceding differences. This difference is due to the fact that the subjects in the Hostility condition listened and talked more to the instigator.

Continuation of Task I. After the instigator left the room and the emotional manipulation was completed, the subjects continued to work for 3 more minutes on the initial task. Hostility and Control subjects performed at virtually the same rate, indicating that any emotional arousal induced by the hostility manipulation did not lead to deterioration of performance on Task I.

Task Changeover. When Task II was introduced, both groups showed a striking decrement in performance. From a response rate of about 156 per minute, the Control group dropped to a rate of 97.8 per minute, and the Hostility group to a rate of 77.0. The difference in response rate between the two arousal conditions is statistically significant ($p = .01$) for the first 3 minutes after changeover. Thereafter both groups improved rapidly and the difference between them was attenuated.

DISCUSSION

It seems clear that the data are in general accord with the predictions. Emotional arousal had no effect on the performance of Task I and a considerable, although transitory, effect on performance after the task changeover. These results serve both to corroborate the findings of the field study by Schachter et al. and to demonstrate their generality to an "artificial" laboratory situation.

This generality becomes the more remarkable when the time span of the two studies is compared. While Schachter et al. prolonged their emotional manipulation over 2-4 weeks, we used 2 minutes; while their subjects had become stereotyped on the same job for untold months, ours had to learn a task in 12 minutes. It is not inconsistent, then, that the emotion-produced decrement on nonstereotyped production which they observed lasted from 1 to 3 weeks after changeover, while we obtained effects for only 3 minutes.

The extreme temporal condensation of our study was not without its disadvantages. The emotional manipulation, while effective, was not particularly strong; and the stereotypy achieved on the first task was problematical at best, considering the continuing linearly improving performance on this task. This latter difficulty is particularly troubling in the light of our prediction that emotional arousal should interfere with nonstereotyped performance. Whether or not performance on Task I can be considered perfectly stereo-

typed, however, it was certainly more nearly stereotyped than Task II performance. Perhaps, then, the most surprising thing about these results is their strength in contrast to the weakness of the manipulations.

REFERENCE

- SCHACHTER, S., WILLERMAN, B., FESTINGER, L., & HYMAN, R. Emotional disruption and industrial productivity. *J. appl. Psychol.*, 1961, **45**, 201-213.

(Received October 1, 1962)

DEPENDENCY RESPONSES TO TELEVISED INSTRUCTION

SHEPARD A. INSEL

San Francisco State College

KURT SCHLESINGER

University of California, Berkeley

AND WILFRED DESROSIERS

Chabot College

This study compared anecdotal responses of 375 college freshmen exemplifying both effective and ineffective instruction in both televised and conventional classroom instruction. Thus, each S gave 4 anecdotes. When treated in the manner of the Flanagan critical incident technique by 4 independent judges, the anecdotes were classified into 2 broad categories: (a) those behaviors describing essentially active initiation by the learner with a focus on self were included in the self-dependent classification, (b) those critical phrases describing essentially passive participation by the learner and focused outside the self were included in the other-dependent classification. The Ss responses to effective televised instruction were significantly more other-dependent. Their responses to effective conventional classroom instruction were significantly self-dependent.

Televised instruction, appearing on the educational scene within the last decade, has been the object of many investigations studying its comparability to conventional instruction. It seems fairly well substantiated that subject matter achievement occurs equally well in one medium as in the other when conventional evaluation methods are used (Carpenter & Greenhill, 1958; Dreher & Beatty, 1958; Erickson & Chausow, 1960; Kumata, 1956; Lepore & Wilson, 1958).

The effects of televised instruction on attitudes and values are somewhat less clear, though it is apparent that televised instruction itself has achieved greater acceptance if one uses the increased number of participants as the criterion (Ford Foundation, 1961). The studies thus far reported in the literature tend to focus on the assessment of the more specific attitude of accepting televised instruction as a mode of education (Allen, 1960). Broader attitudinal investigations which translate themselves into other areas of behavior have been sparse. In one study, no significant differences occurred between televised and conventional classroom instruction when improvement of self-insights and critical thinking were the variables studied (Lepore & Wilson, 1958). Another study reported no significant effects occurring in an overall sense when a controversial program involving

prejudice was presented, though relatively nonethnocentric viewers engaged in less intense name stereotyping as a result (Evans, Wieland, & Moore, 1961).

In a previous study involving the effects of televised instruction on student performance, the senior author assessed the attitudes of students toward the teaching-learning process in both televised and conventional classroom instruction (see Insel, in Lepore & Wilson, 1958) in a manner aimed at eliciting a more covert level of attitudinal set.

By using a modification of Flanagan's critical incident technique (Flanagan, 1954), written anecdotes were collected and classified in terms of what constituted effective and ineffective instruction from the student point of view. While student and teacher reactions toward the teaching-learning process revealed a common core of incidents reflecting effective and ineffective instruction, specific student attitudes varied from one learning context to another. Although something of value was found in both televised and conventional classroom situations, televised instruction was perceived as allowing fewer kinds of satisfactions and dissatisfactions than did conventional instruction.

In reviewing the data, notice was taken of the fact that the student responses tended to fall into two broad categories: those which

described something of value they saw themselves as being able to do or not do, and those which described something of value which existed or was initiated outside themselves. In this sense, references to self-initiated or self-dependent behavior versus non-self-initiated or other-dependent behavior were seen to emerge in the students' attitudes.

This study deals with the relationship found between self-dependent and other-dependent responses elicited to questions about effective and ineffective teaching-learning experiences in both televised and conventional classroom instruction. Specifically, it was hypothesized that televised instruction would elicit more other-dependent responses than conventional classroom instruction.

PROCEDURE

Three hundred and seventy-five college freshmen were used as subjects. These students were enrolled in one television course taking the remainder of their semester's work in conventional classroom courses. The courses in which televised instruction was offered were biology, psychology, basic communication, and creative arts.

After the tenth telecast each student was asked to describe specific and recent incidents which to them exemplified: effective television instruction, ineffective television instruction, effective classroom instruction, and ineffective classroom instruction. Thus, each student returned four written forms, with an anecdote to exemplify a point of view in each of the four major categories listed above.

These statements were next examined by four clinical judges, who extracted the critical phrase (or phrases) from each, in the manner of the Flanagan critical incident technique. It was noted that the universe of critical phrases clustered around six definable subcategories. The critical incidents were then sorted into the six subcategories by the same four judges. The six subcategories, with an example of each, are listed below:

Teacher-initiated behavior. This subcategory consisted of statements which centered around something the teacher said or did, or did not say or do. Primary emphasis was on some action or responsibility of the teacher. A typical example of a phrase extracted from a statement was: "The teacher directed where to look for details."

Student-initiated behavior. This subcategory included statements which emphasized what the student felt or did. Primary emphasis was on the student's personal initiation of behavior and could be differentiated from interactions with other students or with the teacher. Statements such as the following exemplify critical incidents which were included in this subcategory, "I was able to relate subject matter to myself as a person."

Student-teacher interaction. This subcategory was used to include personal interactions between the student and the teacher on a one-to-one basis even though the interaction was in a class situation. A typical phrase included in this subcategory was "When an idea is unclear, the teacher is there to explain."

Class discussion or interaction. In this subcategory were included forms of behavior or expressions of attitude, feelings, and opinions which involved positive or negative interactions with other students. Interactions in which the teacher participated were excluded. A typical phrase in this subcategory was "Class discussions kept the student's attention."

Application of course material or method. Statements which emphasized the nonperson aspects of the teaching-learning process—statements about visual aids, films, reading material, etc.—were included in this subcategory. A typical example is "Effective use of visual aids accompanying lecture."

Physical aspects of the teaching-learning situation. In this category were included statements which emphasized the room, the physical structure of the communication medium, or the perceptual situation. What the student could do with the medium was not included; what the producer could do with the medium was included. A typical phrase which was included in this subcategory was "Class size does not interfere with the course via television."

A seventh subcategory was also developed to take care of statements which were either unresponsive or of so general a nature that they could not be classified in the other subcategories. Sheets which were returned blank were also included in this category. This seventh subcategory was called *Useless*.

After all incidents had been sorted into these seven categories, it was noted that these categories could be further classified as belonging into one of two logical groups. These were: critical incidents describing essentially active initiation of behavior by the learner, and critical incidents describing passive participation by the learner in the behavior initiated by others. These two logical groupings were labeled *self-dependent* and *other-dependent*, respectively.

The statements included in the six subcategories listed above, exclusive of the *Useless* subcategory, were then classified into these two logical groups by independent sorting of three other judges. Since comparisons were to be made between televised and conventional classroom instruction, and since teacher-student interaction is precluded by the medium of television, responses depicting student-teacher interaction were removed from the sample to avoid a biasing effect in favor of the classroom instruction.¹ These constituted 86 responses to ef-

¹ It should be noted that all teacher-student interaction responses were obtained to classroom instruction; also, that these responses belong in the self-dependent category. Since the hypothesis to be tested is that televised instruction leads to other-

TABLE 1
EFFECTIVE INSTRUCTION

Classroom	TV				Total
	Self	Other	Mixed	Useless	
Self	50	131	8	8	197
Other	21	79	6	7	113
Mixed	6	19	2	6	33
Useless	6	15	3	8	32
Total	83	244	19	29	375

Note.— $\chi^2 = 90.44$; $df = 9$; $p = .001$.

fective classroom instruction and 32 for ineffective classroom interaction.

Now that it was possible to locate their attitudinal emphases in terms of their anecdotes, the students were then classified according to the assignment of their responses into 1 of the 16 combinations as shown in Tables 1 and 2. For example, a student whose anecdotal responses to effective televised instruction were classified as essentially self-dependent and whose responses to effective classroom instruction were essentially other-dependent was assigned to two of the categories in Tables 1 and 2, i.e., into Self-TV and Other-Class. Where a student's responses were classified in both the Self and Other for a single medium, such as effective televised instruction, this was considered a mixed response and classified as such.

RESULTS

The total sample of critical phrases which had been sorted into the seven subcategories by the four judges was tested for interjudge agreement using a weighting system by which incidents with higher interjudge agreement

independent behavior, omission of these responses would err in the "conservative" direction.

received greater weight (Lepore & Wilson, 1958, p. 15). Interjudge agreement approached 80% in 65% of the items.

Table 1 shows the number of students, by type of responses, to effective instruction for both conventional classroom and televised instruction. This gives a highly significant chi square of 90.44 ($p < .001$). One hundred and thirty-one of the sample, or 35% of the group, responded to effective televised instruction in an other-dependent manner, while at the same time describing effective conventional classroom instruction by using essentially self-dependent phrases.

Table 2 shows the number of students, by type of responses, to ineffective instruction for both conventional and televised instruction. This gives a chi square of 8.34 which is not significant.

It is interesting to note that the next largest number of students appeared in the category describing other-dependent responses to both televised and conventional classroom

TABLE 2
INEFFECTIVE INSTRUCTION

Classroom	TV				Total
	Self	Other	Mixed	Useless	
Self	16	74	5	7	102
Other	26	100	10	9	145
Mixed	2	16	4	2	24
Useless	11	73	10	10	104
Total	55	263	29	28	375

Note.— $\chi^2 = 8.34$; $df = 9$; $p = ns$.

instruction and that this group represented the largest proportion describing ineffective instruction.

DISCUSSION AND CONCLUSION

That televised instruction evoked, or at least was related to, significantly more other-dependent responses in the students than did conventional classroom instruction seems appropriate. To report what is effective and ineffective in a teaching-learning situation where there is little opportunity to foster communication with others, requires focusing attention on forces outside the learner. There seems to be no choice but to identify external factors within a framework of value if one is to feel some sense of meaning in an experience.

However, it is more difficult to account for the frequency of other-dependent responses when the participants in the teaching-learning process are contained in a common room. It may be that today's conventional classroom instructional practice does less to encourage as much active participation on the part of the learner than might be expected. Consequently, the learner comes to accept more of an other-dependent attitude as being appropriate and satisfying.

Since the students experienced both media, the fact that there was a significant shift in the dependency responses suggests that what constitutes effectiveness in a particular process is to a greater extent a function of context, rather than a value residing characteristically in the individual. If this is so, then for these people, concern that televised instruction is but one more process to reduce the initiative and individuality of the person may be less warranted. Obviously, those who

are characteristically self- or other-dependent are not going to be particularly swayed by the context of the instructional situation.

Nevertheless, unidirectional communication by its very structure invites other-dependency if there are few resources built into the individual learner to check or confirm what he perceives, or if the situation prevents active checking processes by the learner. More data are necessary to assess the long-term effects of these processes.

REFERENCES

- ALLEN, W. H. *Television for California schools*. Sacramento: California State Department of Education, 1960.
- CARPENTER, C. R., & GREENHILL, L. P. *An investigation of closed circuit television for teaching university courses*. University Park: Pennsylvania State University, 1958.
- DREHER, R., & BEATTY, W. *An experimental study of college instruction using broadcast television*. San Francisco: San Francisco State College, 1958.
- ERICKSON, C. G., & CHAUSOW, H. M. *Chicago's TV college: Final report of a three-year experiment*. Chicago: Chicago City Junior College, 1960.
- EVANS, R. I., WIELAND, B., & MOORE, C. W. The effect of experience in telecourses on attitudes toward instruction by television and impact of a controversial television program. *J. appl. Psychol.*, 1961, **45**, 11-15.
- FLANAGAN, J. C. The critical incident technique. *Psychol. Bull.*, 1954, **51**, 327-340.
- FORD FOUNDATION. *Teaching by television*. (2nd ed.) New York: Ford Foundation Office of Reports, 1961.
- KUMATA, H. *An inventory of instructional television research*. Ann Arbor, Mich.: Educational Television and Radio Center, 1956.
- LEPORE, A., & WILSON, J. D. *An experimental study of college instruction using broadcast television: Project number two*. San Francisco: San Francisco State College, 1958.

(Received October 8, 1962)

DEFINING THE PERCEIVED FUNCTIONS OF PURCHASING PERSONNEL

J. C. DENTON

Psychological Business Research, Cleveland, Ohio

AND ERICH P. PRIEN

Western Reserve University

To identify the functions performed by a purchasing division as perceived by company employees using the purchasing service, a 42-item mail-out questionnaire was constructed from intensive interviews. Questionnaire responses were factor analyzed by the centroid method and rotated to approximated simple structure. 8 factors were obtained, namely: buying supplies, equipment and services; protecting the company's capital and assets; "customer" requisitioner activities; optimizing inventory; controlling risks in dealing with vendors; assuring the purchase of standard, high quality commodities; enforcing government regulations; and assuring vendor performance. Factor scores assumed to reflect importance as perceived by respondents showed low correlations with requisition frequency and average dollar value of the requisitions they submit. Factor scores averaged by company departments reflected considerable variation in assumed importance.

The functions of individuals who comprise a business organization depend upon the nature of the organization, the real needs created by the organization structure and the operations performed, and by the perceptions of the individuals within the organization. In turn, both the actual and the perceived performance of the organization and the individuals are dependent upon the same factors with even greater emphasis on perceptual variables. In this context, the development of an effective organization and utilization of the individual members require a concise definition of functional requirements, including those based on perceptions and personal judgments.

This thesis is apparent from another point of view. Consider the ways in which we measure the success of a business enterprise—usually in terms of total sales, dollar profits, dividends, or other financial indicators. However, the contribution to this success made by some segments within the business is difficult to measure. Legal sections or employee relations departments encounter problems in showing where they help to create profits. Yet they provide services and fill needs equally as well as does manufacturing or sales. What is true for a staff section is also true for a given individual—it is not

easy to tell which man is effective. To measure an individual's performance, we need a clear-cut description of the functions he is to perform in his position. Measuring the relative performance of any unit in a dynamic personnel system is a first step in generating the strategies that would increase the effectiveness of the whole enterprise.

Because its basic mission is to provide services which fill the needs of customers,¹ the Purchasing Division of a large midwestern manufacturer elected to study its customers with the objective of determining what functions these customers *thought* Purchasing performed, and the relative importance customers attached to these perceived functions. The "real" functions in the case of purchasing are fairly clearly defined and documented through the established requisitioning procedures.

METHOD

Questionnaire Construction

The basic approach was to develop a questionnaire that could be mailed to selected personnel in

¹ The term "customers" is used here to identify company employees who requisition supplies and services from and deal with personnel of the Purchasing Division. In a real sense, they are customers in that a service is obtained and the individual has some discretion in how the service is used.

the company who had some contact, direct or indirect, with the Purchasing Division or function. In order to develop this survey form, interviews were first held in the offices of the company with a small sample of purchasing personnel. This was followed by interviews with 22 people in other departments. The interviewing was unstructured, exploring areas of concern to the respondents. The objective was to obtain relevant ideas for inclusion in the mail survey. Out of this pool of information, 42 statements were drawn which reflected the most significant aspects of the interviews. An attempt was made to preserve the "working language" of the respondent—not to insert the more precise language that might be appropriate for policy or procedure manuals.

The directions for the questionnaire required the respondent to judge whether or not a given activity, such as "Help dispose of excess inventory" was something that Purchasing did. If it was not an activity they performed, he marked a zero next to the statement. On the other hand, if he believed Purchasing did perform this activity, he was asked to judge the relative importance of this activity on a 5-point scale. It should be noted that the questionnaire asked customers of Purchasing to tell what they thought Purchasing did. In a very real sense this is a marketing study, the purpose of which is to adapt ultimately the operations of an organization to meet more effectively the needs and perceptions of its customers.

Questionnaire Administration

The questionnaire was pretested with a small sample of employees. Surveys were then mailed to a stratified sample of personnel in the company. About 10% of the total corporation personnel were given forms to fill out. In order to provide meaningful results for the Purchasing Division it was necessary to learn something about the respondent. The survey form asked each one to indicate his department by checking an appropriate blank space. In addition to the 42-item responses, data as to the frequency of respondent contact with Purchasing was included. Also an estimate was made of the average dollar value of the respondent's typical requisition sent to Purchasing.

Analysis

The 42 items were intercorrelated and the matrix was factor analyzed by the centroid method. Eight factors were obtained and were rotated maintaining orthogonality to an approximate simple structure. A graphic plotting of the machine rotated factors indicated only a minor adjustment to be made in the last factor extracted.

RESULTS

While 637 (72.4%) questionnaires were returned, not all of them were usable. For instance, some of the customers did not feel

sufficiently close to the activities of the Purchasing Division to provide meaningful information. The final results are based on factor analysis of 527 cases.

Factors

The organization of the results presented here follows a pattern. First a factor will be named and defined. Next the items² which are most significantly descriptive of the factor are listed in order of their importance (or contribution) to that factor. It must be remembered that this is a factor analysis of perceptions of typical customers of a service organization. Factors comprised of statements about *activities* of buyers are best considered as *functions*. Thus "To obtain comparative bids on some equipment" would be a daily activity of a buyer. Several such activities combine into a cluster because the customer perceives them as having common elements—he perceives these as the functions of the buyer.

Factor I: Buying supplies, equipment, and services. The major emphasis of all activities in this factor is upon dealing with the vendors in order to obtain favorable results for the company with respect to prices, delivery, and quality of the commodity. The concern is with maximizing the financial gain for the company. The items that are most significantly descriptive of this factor are:

15. Determine the supplier of equipment, etc.
23. Obtain comparative estimates and bids on all sizable purchases
27. Determine which supplier is to get an order

In the analysis of departmental results, a high score would indicate that customers in the particular department believed these kinds of activities to be of considerable importance to the operation of the department and required the careful attention of the buyer. To them, buying is an important aspect of Purchasing's activities.

Factor II: Protecting the company's capital and assets. The activities within this factor are largely concerned with prices, credits, and

² Item number in the questionnaire corresponds to item numbers in Table 2.

TABLE 1
ROTATED FACTOR LOADINGS

Item	Factor loading								<i>h</i> ²
	I	II	III	IV	V	VI	VII	VIII	
1	01	07	−54	07	−03	04	−04	−02	31
2	13	−08	−13	−22	−17	−18	−14	17	20
3	00	−17	−13	07	−10	06	−39	02	22
4	03	−16	−43	10	00	21	−10	10	29
5	14	04	−38	−05	−09	03	−25	−12	25
6	55	−10	−04	−06	−27	−02	−02	29	48
7	26	−24	−06	−06	−28	−13	−35	02	35
8	40	−17	−03	−26	−12	−04	−28	09	36
9	34	−07	−17	−35	00	06	−10	−03	29
10	37	−08	−08	−23	−19	04	−24	−18	33
11	27	−31	−18	−48	03	00	06	00	44
12	08	−24	−46	−34	−06	−01	11	00	41
13	29	−40	−13	−18	−19	02	12	16	37
14	22	−47	02	−26	14	−10	07	02	37
15	61	16	−17	10	14	−07	−02	02	46
16	21	−52	−16	05	13	−02	−11	15	39
17	59	−25	−04	−10	−17	04	09	32	56
18	36	−21	−13	−26	−10	−16	−07	−07	30
19	21	−14	−47	00	−05	−32	08	00	40
20	39	−13	−28	−02	−03	−40	−01	−06	41
21	28	−44	16	−15	22	05	−14	23	44
22	27	−51	07	−14	07	−07	−06	16	40
23	58	−12	02	−10	00	11	07	−16	40
24	34	−40	00	−13	03	15	06	−14	34
25	48	−25	−11	21	04	−35	−02	09	48
26	45	−38	06	−13	−09	−04	−06	−22	43
27	69	05	02	07	08	−07	−05	01	50
28	39	−30	−10	−10	−06	−06	−08	−27	35
29	63	−15	06	−05	00	15	00	−22	50
30	40	−42	−07	−17	00	10	05	−25	44
31	27	−53	05	−05	01	−09	04	−09	38
32	26	−48	−19	13	−07	−12	06	−05	38
33	11	−52	−09	03	05	05	−08	−17	33
34	09	−52	−06	−18	12	−02	07	02	33
35	03	−21	−52	04	11	10	−01	02	34
36	14	−45	−09	−02	13	−09	−05	06	26
37	05	−56	−03	06	−09	05	−14	−06	35
38	23	−41	02	−07	−47	13	04	−04	47
39	22	−43	−04	06	−31	−21	−11	05	39
40	11	−46	00	−09	−49	−05	02	−02	48
41	17	−57	05	09	−10	13	03	04	39
42	13	−44	−35	16	00	−16	15	04	41

Note.—*N* = 527.

allowances, and involve the assurance that the company does not dilute its efforts but rather optimizes inventories, vender performance, and investment of money. They also involve administrative procedures of the Purchasing Division to control the buying function, minimize risks, and protect the Company from loss in general. The items most descriptive of this factor are:

- 31. Supply accurate price information to the field
- 33. Determine the most economical methods for shipping
- 34. Facilitate the disposal of excess inventory

Factor III: Customer activities. A number of activities were included in the survey which

by management policy are not performed by the Division. These activities were significantly related to the third factor and are further characterized, for the most part, as responsibilities of other departments. Such activities often result in a requisition being sent to the Purchasing Division. The items defining this factor are:

- 1. Determine the need for purchase of new equipment or parts
- 4. Verify quantity and/or quality received from suppliers
- 12. Coordinate inventories and requisition quantities

Some of these activities, regardless of their appearance on this factor, are in fact a con-

TABLE 2
UNROTATED FACTOR LOADING

Item	Factor loading								h ²
	I	II	III	IV	V	VI	VII	VIII	
1	11	-10	44	14	-23	03	-13	-05	31
2	26	-12	06	23	13	12	14	09	20
3	18	05	20	14	09	-19	17	-22	22
4	23	10	33	15	-13	03	-15	-25	30
5	22	-21	27	17	-19	-12	06	-10	26
6	48	-28	-12	15	28	13	-18	-07	48
7	43	-11	06	17	22	-19	18	-02	35
8	47	-21	-12	17	08	03	20	-09	36
9	38	-20	-11	16	-21	11	11	-05	29
10	40	-26	-11	17	-10	-19	13	-04	33
11	50	04	-16	16	-24	22	12	11	44
12	39	08	15	25	-30	19	-03	17	40
13	53	11	-10	15	06	13	-13	08	37
14	47	20	-18	-12	-05	18	14	09	37
15	31	-50	05	-24	-03	13	-13	-15	46
16	51	24	12	-13	09	11	06	-15	40
17	58	-16	-19	08	22	22	-23	-09	56
18	49	-16	-05	06	-05	02	12	14	31
19	38	-12	38	-02	-07	12	-07	25	39
20	46	-27	22	-14	03	07	05	24	42
21	43	19	-23	-13	13	21	19	-24	45
22	51	20	-15	-09	14	14	14	-05	39
23	45	-24	-27	-14	-13	-06	-14	-06	40
24	48	11	-22	-09	-14	-06	-05	-05	33
25	49	-18	16	-31	26	08	-04	07	47
26	58	-05	-22	-11	-02	-18	07	06	44
27	39	-47	-12	-27	08	05	-09	-15	50
28	52	-09	-06	-09	-13	-19	06	06	35
29	49	-26	-30	-18	-13	-14	-12	-13	50
30	58	05	-20	-10	-21	-12	-03	02	45
31	52	20	-13	-18	05	-05	04	09	37
32	53	15	14	-15	07	-08	-11	09	38
33	43	29	03	-12	-08	-16	06	-05	32
34	43	33	-08	-06	-07	13	09	06	34
35	28	12	38	04	-26	10	-10	-11	34
36	41	23	06	-14	04	09	10	-02	26
37	43	35	04	-02	08	-20	06	-06	36
38	47	13	-19	25	14	-25	-19	09	46
39	49	08	07	05	31	-15	02	13	39
40	46	20	-10	26	21	-22	-09	22	48
41	46	34	-11	-06	13	-11	-12	-08	39
42	45	21	31	-12	12	07	-17	13	42

Note.—N = 527.

cern of the buyers, and it was considered important to include in the final analysis another item for activity that showed a low relationship to this factor:

20. Help write specifications to promote standardization

This factor represents an area of activity where careful definition of authority and responsibility is needed for Purchasing to work effectively with its customers. For this factor there was less agreement across segments of the company as to the relative importance of the activity to Purchasing.

Factor IV: Optimizing inventory. The emphasis here is upon the efforts of Purchasing to cooperate with the operating departments in the mutual responsibility of maintaining adequate inventories without investment of excess capital. The items that define this factor are:

- 9. Determine the best timing for purchases
- 11. Suggest improvements for stock or repeat items in quantity
- 12. Coordinate inventories and requisition quantities

Two other items might have been judged pertinent to this function, but the factor analysis showed they were not. These items are:

- 14. Help assure adequate trade-in allowance on salvage or surplus items
- 34. Facilitate the disposal of excess inventory

While both of these items are somewhat related to this factor, the first has a higher loading on Factor I while the second has a higher loading on Factor II.

Factor V: Controlling risks in dealing with vendors. There are a number of activities of Purchasing where the major objective appears to be concerned with assuring vendor performance by minimizing insecure situations. Purchasing attempts to determine the probability that a vendor has a sound operation and will be able to perform without jeopardizing the company. The items most descriptive of this factor are:

- 38. Obtain Dun & Bradstreet reports on potential suppliers
- 39. Handle confidential material (viz., patents) with special care
- 40. Keep track of supplier's business problems and trends

It should be noted that this factor is not concerned with positive action taken to get a vendor to perform as might be the case in "punching up" a vendor, but one preventive and usually performed prior to the placing of an order or a contract.

Factor VI: Assuring the purchase of standard, high quality commodities. This factor represents the perceived activities of Purchasing undertaken in order to determine if the commodity meets specifications and standards for quality and performance. The most significant items are:

- 19. Conduct tests to define the quality and performance of equipment, parts, etc.
- 20. Help write specifications to promote standardization
- 25. Check specifications versus a supplier's offering on equipment

It is significant to note that the general character of the above items is anticipatory in the same sense that Factor V is preventive. In other words, these items do not reflect a concern with checks made after the order has been shipped. "Verifying the quantity and/or quality received from suppliers" (Item 4) and "Point out where local purchases are out of line on price" (Item 24) are two activities that are *not* related to this factor, although they might be postulated to have a relationship. Both of these are "control" not "budget" oriented.

Factor VII: Enforcing government regulations. There are three items that clearly define this factor. They are:

- 3. Control the tax status of material ordered
- 7. Avoid difficulties with federal regulation, e.g., price fixing
- 8. Promote good trade relations

The loading of Item 8 on both Factors VI and VII provides some interesting material. It seems that customers expect buyers

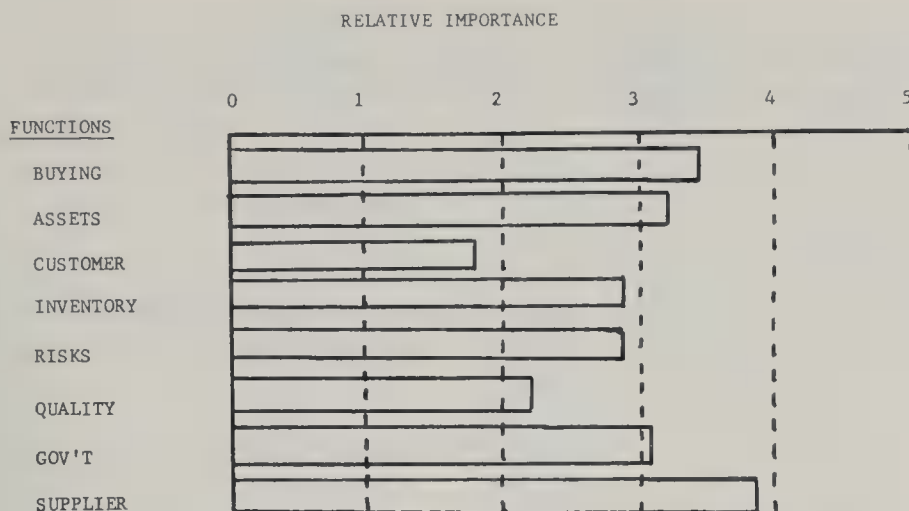


FIG. 1. Mean factor scores for total sample of customer respondents.

to maintain a friendly, cooperative relationship with a vendor in order to assure the receipt of high quality products. On the other hand, this relationship must not be so friendly or cooperative as to bring about the judgment of collaboration or collusion by the federal government. Some sort of an "optimal" relationship must be established with the vendor.

Factor VIII: Assuring vendor performance. There are many activities which the Purchasing Division performs that are primarily centered around completion of contracts and assuring the delivery of the commodities ordered. They are responsible to see that the vendor does in fact perform satisfactorily—that he is now operating efficiently. This is a control relationship with the vendor where price is not the primary concern of the buyer—rather the performance of the vendor in doing his job. Items which are significantly related to this factor are:

6. Check a supplier for his efficiency and ability to do the job
17. Analyze the performance of suppliers
21. "Punch up" suppliers as needed to expedite delivery

It is possible to portray all the findings of this survey in chart form. Thus, the average factor score of all personnel in the company can be shown for each of the factors found (see Figure 1). The "score" here has been expressed on a scale corresponding to the way the customer first checked the statements in

the survey, i.e., on the 1–5 scale. It was also possible to obtain the average factor score for any segment of the company, such as Manufacturing Department A.³

In addition, factor scores were computed for individual respondents and correlated with the frequency and average dollar value of requisitions by the individual. This was done on the assumption that these indices are criteria of the intensity and extent of contact between the customer and the Purchasing Division. Obtained correlations were low, reaching a maximum of .20 with scores on the Buying factor as was anticipated.

DISCUSSION

The results of this study can most realistically be considered as a management report, as a tool or device for management to use in operating the business. The results reflect the current status of one aspect of the organization, the perceived functions performed by the Purchasing Division and the relative importance of each function to the other departments in the company.

As one effort to answer a question of relative importance, factor score distributions were obtained by departments of the company for each of the eight factors. The variability of the distribution is a measure of the degree

³ The specific results, since they are unique to the problems and needs of each department and certainly unique to the company as a whole, are not presented in this report.

of agreement or disagreement across the company concerning the "importance" of a factor. As expected, the variance, factor by factor, differed markedly. Most customers in the Company were in good agreement as to the relative importance of the buying and preservation of assets, and to a lesser degree, of insuring vendor performance. At the other extreme—that of disagreement—are the factors of optimizing inventory, of controlling risks in dealing with vendors, and of enforcing government regulations. While consistency of judgments made by the customers may not be the best indicator of importance, inconsistency of agreement might imply that Purchasing has a communication problem to meet in future relationships with segments of the company.

It is not possible with a survey of this kind to generate a scale of importance to attach to each factor which would actually show the contribution of that factor to the profit picture of the company. As is the case with many phenomena that are dependent upon human behavior, the importance of something is often dictated by the perceptions of people as opposed to dollar figures, or other finite indicators. The factors obtained can be the basis around which many personnel controls and administrative activities are centered. For example, the factors themselves can be considered as bases for preparing a description

of a buyer's job. As such, some new concepts have emerged, and with rather more crisp definition than might have been possible with the usual job analysis techniques. It is unlikely that any buyer's job has been heretofore defined carefully as including the performance of eight different functions.

Secondly, the activities defining the factors become the core about which a complete system of performance standards can be developed. Such performance standards would be meaningful, well defined, and quite descriptive of actual buyer performance.

A third extension of the results would be to use them as part of an orientation seminar for new Purchasing employees, or to organize in a new employee form more meaningful concepts with respect to his activities and responsibilities. While there are almost infinite potential applications, the primary point of consideration is the fact that individual respondents as customers and specific departments differ as to the importance they attach to the activities of the Purchasing Department. While conformity behavior as such may not be a socially promotable concept, consistency of emphasis, particularly when selective, between the purveyor and recipient of a service seems in itself to be a desirable end.

(Received October 15, 1962)

LEADER ASSUMED DISSIMILARITY AS A MEASURE OF PREJUDICIAL COGNITIVE STYLE¹

ROBERT C. ZILLER

Center for Research on Social Behavior, University of Delaware

In a series of studies Fiedler concluded that leaders who maintain a greater psychological distance between themselves and the group members are more effective in promoting group productivity. Statistical and sampling shortcomings of these studies led to the present study involving 43 infantry teams in training. The leaders of high teams as compared to low teams (as rated by their commanding officers) evaluated a least and most preferred co-worker on a series of dyadic adjectival scales. Leaders of high rated teams evaluated the low reference person higher ($p < .05$) and used the lowest evaluation point less frequently when describing the least preferred co-worker ($p < .05$). The results suggest that under conditions where the leader is not an agent of selection, leaders of more successful teams are less severe in their evaluations of members with lower achievement potential.

Fiedler (1960) has summarized the results of an impressive series of studies indicating that leaders who maintain a greater psychological distance between themselves and the members (assumed dissimilarity) are more effective in promoting group productivity than are leaders with psychologically closer interpersonal relations. By way of explanation it was proposed that psychologically closer and warmer relations with subordinates make it difficult for the leader to impose disciplinary measures and encourage rivalries and charges of favoritism.

Although a variety of groups in a variety of settings were included in the aforementioned series of studies, closer analysis of these groups suggested that the sample was less catholic than had heretofore been assumed. In the studies of basketball teams, surveying teams, B-29 bomber crews, and open hearth shops, the groups involved were in an advanced stage of development. All groups were well beyond the initial stages of training; all were being evaluated with respect to their ability to achieve a high standard of performance according to objective and patent criterion. In addition, the leaders, in most cases, were in positions which enabled them to control the selection,

assignment, and/or replacement of personnel. Finally, the results pertain, for the most part, to groups in which the leaders were accepted by the group members. These restrictions on the experimental groups necessarily restrict generalizations of Fiedler's findings.

Still, the major criticism of the program of research concerning assumed similarity is of a statistical nature (Cronbach, 1958). The index of assumed similarity is essentially the difference between the descriptive ratings of a person with whom the respondent has been able to work best and the person with whom the respondent has been able to work least well. Cronbach (1958) incisively observed that this subtraction score "leads to an unparsimonious description of events" since the descriptions of the high and low persons could be compared separately, and that "dyadic analysis is a breeding ground for artifacts."

In answer to these criticisms, several more direct methods of analysis of the same basic data have been suggested (Cronbach, 1958). Thus, Kirchner (1961) and Hawkins (1962) both demonstrated that better supervisors (as rated by their superiors) showed less leniency in their ratings of subordinates, but especially toward the low reference person in the Fiedler dyadic scale forms. Both of these latter experiments suggest that the leaders of the more productive industrial units are

¹ This investigation was supported under Contract Number DA-49-083 OSA-2321 from the Research and Development Division, The Adjutant General's Office, Department of the Army.

more punitive with regard to a subordinate whose performance is marginal or less.

The present study was undertaken in an effort to investigate the applicability of Fiedler's findings to United States infantry squads and artillery sections under training conditions. In this study the leaders of high and low productivity teams, as rated by their commanding officers, were compared with regard to assumed similarity indices; mean ratings of the low reference co-worker; mean ratings of the high reference co-worker; and points selected on the descriptive ratings scales. The latter analysis provides a more detailed description of the leader's rating tendencies than any attempted heretofore.

PROCEDURE

Subjects

Forty-three United States Army teams from Fort Benning, Georgia, participated in the study. The 43 teams included 11 howitzer gun sections and 32 infantry squads along with their leaders who were noncommissioned officers. The teams varied in size from 5 to 10 members. A team of median size was composed of a noncommissioned officer and 6 enlisted men. No team was composed of less than a leader and 5 men or more than a leader and 10 men. A total of 300 soldiers was involved in the field study.

The measure of team productivity was an overall rating of the teams' military effectiveness submitted jointly by the teams' platoon leader and the commanding officer of the infantry company. The ratings were submitted on an 11-point scale ranging from "one of the worst teams I have seen" through "an average team" to "one of the best teams I have seen." Only these 3 points on the scale were described.

The test format used contained 20 sets of personality adjectives and their antonyms with each pair separated by a 6-point scale (Fiedler, 1960, p. 590). Some of the adjectives included friendly-unfriendly, cooperative-uncooperative, quits easily-keeps trying, confident-unsure, bold-timid, and careless-careful. The subject is given two identical scale sheets and is asked to describe the "person with whom you can work best" and the "person with whom you can work least well." His most and least preferred co-workers may be individuals with whom he currently works, or they may be people he has known in the past. To obtain the measure of assumed dissimilarity, corresponding items on the two scale sheets are compared and the differences between scores on corresponding item pairs are squared and summed. The square root of the squared and summed differences provides a

reasonably normal distribution of scores (Fiedler, 1960).

RESULTS

The correlation between the leader's assumed similarity score and the team's rating of overall effectiveness was .22 ($N = 40$). These results are opposite to those reported by Fiedler. Here, the leaders who evaluate the high and low reference persons *more similarly* (low psychological distance in Fiedler's terms) are in command of *more productive* teams.

In a second analysis, the teams were divided into high and low effective units. This was accomplished by dividing the team ratings at the point where an even number and maximum number of teams were clearly differentiated. In this manner 17 high and 17 low rated teams were identified and compared with regard to the leaders' mean rating of the least preferred co-worker and most preferred co-worker (see Table 1). There was no statistically significant difference between the leaders of high as opposed to low rated teams with regard to their evaluations of the high reference persons in the dyadic scales; but the leaders of the high rated teams were found to rate the low reference person *higher* ($t = 2.18$, $df = 32$, $p < .05$).

The final analysis concerns the specific rating scale points used by leaders of high rated teams as compared with leaders of low rated teams in describing the least preferred co-worker. In a 2×6 analysis of variance for repeated measures (Edwards, 1960) involving the leaders of high and low rated teams and 6 rating scale points, the interaction effect between high and low leaders and scale points was statistically significant.

TABLE 1
MEAN EVALUATIONS OF LEAST PREFERRED AND MOST PREFERRED CO-WORKERS BY LEADERS OF HIGH AND LOW RATED MILITARY TEAMS

Co-workers	High rated teams	Low rated teams	<i>t</i>
Most preferred	4.71	4.76	.22
Least preferred	3.69	2.90	2.18*

Note.— $N = 34$.
* $p < .05$.

($p < .05$) (see Tables 2 and 3). The high leaders differed from low leaders in the frequency with which they used the 6 rating scale points with reference to the least preferred co-worker. Inspection of the data revealed and the Tukey HSO test (Ryan, 1960) confirmed ($p < .05$) that the leaders of the high rated teams used the lowest evaluation scale point less often than the leaders of low rated teams.

DISCUSSION

The results are opposite to those reported in earlier investigations involving dyadic rating scales. Previous research by Fiedler (1960) demonstrated that leaders of more productive groups maintained a greater psychological distance between themselves and the group members. The results of the present study indicate that the leaders of the higher rated teams maintained less psychological distance between themselves and the group members. The results of two previous studies involving industrial groups (Hawkins, 1962; Kirchner, 1961) and a study involving farm cooperatives (Cronbach, 1958) revealed that leaders of more productive groups were more severe in their ratings of the least preferred co-workers than were leaders of less productive groups. In the present study, the results were diametrically opposite. Finally, in the

TABLE 2

MEAN PROPORTION (arc-sine transformation) OF THE HIGH AND LOW LEADER'S USE OF THE LEAST PREFERRED CO-WORKER SCALE

Mean use of scale points		
Scale points ^a	Leaders of high rated teams	Leaders of low rated teams
6	19.0	18.9
5	23.4	14.6
4	19.1	14.8
3	24.0	19.4
2	16.7	15.8
1	14.8	35.3

Note.—Since all leaders did not make the same number of ratings, the proportion of each leader's use of individual scale points was calculated and the arc-sine transformation of these scores constituted the dependent measure.

^a A low scale point represents a less desirable description.

TABLE 3

ANALYSIS OF VARIANCE INVOLVING LEADERS OF HIGH AND LOW RATED TEAMS CONCERNING THE LEADER'S USE OF THE SIX POINTS OF THE LEAST PREFERRED CO-WORKER SCALE

Source	MS	df	F
High-low leaders (A)	5.18	1	—
Scale points (B)	360.24	5	1.08
A × B	918.15	5	2.75*
Error (C, A) ^a	58.53	32	—
Error (BC, A) ^b	334.04	160	—

^a Represents the error variance among leader's ratings across scale points.

^b Represents the error variance among leader's ratings at each scale point.

* $p < .05$.

present study it was found that the leaders of low rated teams used the lowest rating scale point when describing the least preferred co-worker more often than did the leaders of the high rated teams. An explanation of the contradictory results was sought by re-examining the characteristics of infantry teams and the objectives of the team leaders.

The first line supervisor of the military training unit is presented with a heterogeneous collection of recently inducted soldiers varying widely with regard to physical, psychological, and sociological characteristics. The group members do not select the group, and the group leaders in no way are involved in the selection of the members. Moreover, since service in the United States Army is compulsory for the conscripted members and, indeed, for all members of the armed forces who have taken the service oath, poor performance is insufficient cause for dismissal and the leader is constrained to utilize the given human resources optimally. Under these conditions, one of the primary functions of the leader of an infantry training unit is to facilitate the rapid development of all the members of his unit toward a basic level of performance. In this process, the leaders necessarily are most concerned with the least preferred team members whose marginal performance threatens to immobilize or seriously retard the group's development and overall performance. Thus, in contrast to groups in earlier studies of assumed similarity, the

leaders in the present study were assumed to be charged with the task of the optimal development of all the members rather than with the selective screening of the superior members.

Against this background, the results suggest that the leaders of higher rated training teams show more concern for and encourage the development of the members whose performance is marginal. Furthermore, the results suggest that the leader is most successful in working with these less effective team members if the leader does not perceive, categorize, and condemn the less talented or less motivated members as untrainables or incorrigibles. This positive attitude may be attributed, in part, to greater cognitive complexity or to less narrowly and less rigidly defined standards of acceptance within an evaluation system or style that includes more than dichotomized rating scales such as good-bad and trainable-untrainable.

Similar results to these may be expected with regard to similar groups. Thus, it is hypothesized that in any teams compelled by economic or labor conditions to accept and utilize a high percentage of job applicants, the leaders of the more productive teams will be found to possess similar perceptual proclivities with regard to the less desirable

employees. Similarly, it is hypothesized that in the early elementary school grades and particularly in schools with a high percentage of underprivileged children, the most successful teachers (in terms of student performance at given minimum standards of achievement by the maximum number of students) are less severe in their adjectival descriptions of the least preferred student.

REFERENCES

- CRONBACH, L. J. Proposals leading to analytic treatment of social perception scores. In R. Tagiuri & L. Petrullo (Eds.), *Person perception and interpersonal behavior*. Stanford: Stanford University Press, 1958. Pp. 353-380.
- EDWARDS, A. L. *Experimental design in psychological research*. New York: Reinhardt, 1960.
- FIEDLER, F. E. The leader's psychological distance and group effectiveness. In Cartwright & Zander (Eds.), *Group dynamics*. New York: Row, Peterson, 1960. Pp. 586-606.
- HAWKINS, C. A study of factors mediating a relationship between leader rating behavior and group productivity. Unpublished doctoral dissertation, University of Minnesota, 1962.
- KIRCHNER, W. K. Differences between better and less effective supervisors in appraisal of their subordinates. *Amer. Psychologist*, 1961, 16, 432-433. (Abstract)
- RYAN, T. A. Significance tests for multiple comparisons of proportions, variance, and other statistics. *Psychol. Bull.*, 1960, 57, 318-328.

(Received October 22, 1962)

ACTIVE RESPONDING IN PROGRAMED LEARNING MATERIALS

THOMAS F. HARTMAN,¹ BARBARA A. MORRISON, AND MARGARET E. CARLSON

International Business Machines Research Center, Yorktown, New York

This study investigated the effects of requiring a constructed response in programed material versus reading the same material. The programed materials consisted of 4 sections of an IBM basic machine operation course. 23 intact classes at the IBM New York Education Center served as Ss. The principal findings were as follows: no significant differences in mean achievement between the 2 formats of programed instruction, some tendency for the brighter students to do better with the constructed response programs and the duller students to do better with reading only, large time savings when the constructed response was not required, and significant reduction in range of finishing times when the constructed response was not required.

One important property of teaching programs according to Skinner (1958) is that the student is required to make an overt response. Since the importance of overt responding is easily investigated in simple programed texts, much research has been devoted to this property. Many gradations of overtness of response have been investigated, from writing the response in the booklet for every frame at one end of the continuum, through "optional responding" (writing in only the answers the subject is sure are correct), "thinking response" (subject thinks of answer rather than writing it), to the other end of the continuum, reading the sequenced program material as text with no response requirement. Results of studies utilizing some of the above response modes are inconclusive; several studies (Campbell, 1961; Coulson & Silberman, 1961; Goldbeck, Campbell, & Llewellyn, 1960; Roe, Massey, Weltman, & Leeds, 1960) reported no differences; Holland (1960) reported "good" programed text superior to reading the same material, while Silverman and Alter (1960) found reading superior to constructing the response. Goldbeck (1960), to further complicate matters, reported an interaction between response requirements and program difficulty. All the studies cited can be criticized for using extremely short programs, small numbers of subjects, or both.

METHOD

Materials. The programed texts used were four sections of a 3,500-frame course on basic machine operation developed at the IBM Research Center. Section I detailed the general characteristics of the IBM punched card (251 frames), Section II described the operation of IBM card punches (382 frames), Section III described the operation of IBM sorters (441 frames), and Section IV described the operation of IBM reproducers (682 frames). The sections differed somewhat in internal organization. Section I contained details about the punched card that were not to any large degree sequentially dependent upon each other, Section II dealt with a more unified single sequence organized about the passage of a card through the card punch, and Sections III and IV described several alternative sequences of operation that could be accomplished on the sorters and reproducers.

Each program section was published in two editions. Edition 1 consisted of varying numbers of frames arranged vertically on each page with frames printed on both sides of the pages. The answers to the blanks in the frames were located in a section at the rear of the booklet. Edition 2 was produced by typing the answers in the blanks in Edition 1 and reproducing the resulting copy, so that the responses to be constructed in Edition 1 were underlined in Edition 2.

Subjects. The subjects were 493 customer trainees in 23 classes of the basic machine operation course at the IBM New York Education Center between December 1, 1961 and September 1, 1962. Sixteen classes received one program section and seven classes received two program sections. Assignment of an edition of the program to a class within program section was essentially random. Only subjects with a score of 15 or above on the Punched Card Machine Operator's Aptitude Test (PCMOAT) were included in the analyses as it was found that many subjects with lower scores experienced dif-

¹ Now at Pennsylvania State University.

TABLE 1
PREDICTED PERCENTAGE QUIZ SCORES ON FOUR PROGRAMED SECTIONS FOR STUDENTS WITH
PCMOAT SCORES SELECTED FROM THE RANGE OF PCMOAT SCORES OBTAINED

Program section	Condition	PCMOAT score			
		15	25	35	45
Punched cards	Constructed response	73.7	81.2	88.7	96.2
	Reading only	77.1	84.1	91.1	98.1
	Constructed response	78.0	85.6	93.2	100.7
Card punches	Reading only	83.6	88.1	92.5	97.0
	Constructed response	59.2	70.8	82.4	94.0
Sorters	Reading only	75.4	79.7	84.0	88.3
	Constructed response	50.3	64.4	78.4	92.5
Reproducers	Reading only	60.3	68.4	76.5	84.6

ficulty in reading and writing the English language. This truncation resulted in a loss of approximately the lower 10% of the population.

Procedure. A member of the research staff distributed the programed texts and read two pages of instructions to the subjects. A suggested time for completion of the materials was given to the subjects and they began working on the booklets. Immediately upon completion of the booklet a quiz was given to the subject. The quiz contained from 30 to 48 items, depending on the program section. This quiz was prepared independently of the program from the same course outline that provided the basis for inclusion of material in the programed booklets. Each edition of Section I was given to three classes at the IBM New York Education Center and each edition of Sections II, III, and IV was given to four classes.

RESULTS

To determine if difficulty of the program sections interacted with requiring a written response, the difficulty of the program sections had first to be ascertained; this was done by tabulating the percentage of incorrect or omitted responses for Edition 1 of each program section. Differences among the sections were small; the mean error rates for subjects scoring 15 or above on the PCMOAT were from 4% to 6% for the various program sections. The ordering of the program sections from most difficult to least difficult was as follows: punched card, reproducing punches, sorters, card punches.

Quiz Scores. Analysis of covariance was used to obtain the most sensitive test of

differences in performance on the criterion quiz. The predictor variable was PCMOAT score; the dependent variable was percentage score on the criterion quiz. Despite the reduction of variances which resulted from omitting data of the subjects who scored less than 15 on the PCMOAT, the Pearson *r*'s between PCMOAT score and criterion test score were sizable, ranging from .28 to .68 for the various sections with a mean *r* = .49 (*z* transform). Thus the use of covariance analysis resulted in approximately a 25% decrease in the error variance estimate. All correlations were significantly different from zero at the .01 level.

Table 1 summarizes the performance of the subjects on the criterion quiz. The entry in each cell of this table is essentially the mean percentage for subjects with particular PCMOAT scores, and was estimated within each program edition and section by the least squares linear regression of percentage quiz score on PCMOAT score. Presenting the data in this manner is more desirable than the usual practice of reporting the covariance adjusted means since subsequent analysis disclosed a tendency toward heterogeneity of regression between some experimental groups. It is evident in Table 1 that differences in *mean* performance between the two editions of a program section were small: on the punched card program scores on Edition 1 were approximately 3 percentage points lower

TABLE 2

ANALYSES OF COVARIANCE OF PERCENTAGE SCORES ON QUIZ FOLLOWING PROGRAMED TEXTS

Source	Punched cards		Card punches		Sorters		Reproducers	
	<i>df</i>	<i>F</i>	<i>df</i>	<i>F</i>	<i>df</i>	<i>F</i>	<i>df</i>	<i>F</i>
Constructed response versus reading	1	1.07	1	0.01	1	3.17	1	0.16
Error (<i>MS</i>)	(100)	(139.67)	(164)	(98.14)	(164)	(143.14)	(139)	(224.17)
Heterogeneity of regression	1	0.03	1	2.63	1	11.65**	1	3.33
Within groups regression (<i>MS</i>)	99	(141.04)	163	(97.18)	163	(134.40)	138	(220.47)

** $p = .01$.

than on Edition 2; on the other three program sections the differences in mean performance on the two editions were likewise small. Inspection of Table 1 indicates that for these three program sections the subjects with lower PCMOAT scores did somewhat better when they only read the materials, while subjects with higher PCMOAT scores appeared to do better when they constructed a response to the materials.

The analyses of covariance of percentage quiz scores on the four program sections are summarized in Table 2. The nonsignificant *F*s for Constructed response versus Reading indicate that there were no reliable differences in the mean performance on the criterion quiz between the subjects that constructed a response and the subjects that merely

read the material. The tests for heterogeneity of regression disclosed significant heterogeneity of regression only for the sorter program. Although the slopes of the regression lines were significantly different only for the sorter program section, it is evident in Table 1 that the direction of the difference between the slopes was the same for all program sections.

Time Scores. The Pearson *r*'s between PCMOAT scores and time to complete the program sections ranged from $-.07$ to $-.50$ for the various program sections with a mean $r = -.36$ (*z* transform). The correlations were all significantly different from zero at the .05 level, except that for the sorter program when a constructed response was not required.

TABLE 3

PREDICTED TIME IN MINUTES TO COMPLETE THE FOUR PROGRAMED SECTIONS FOR STUDENTS WITH PCMOAT SCORES SELECTED FROM THE RANGE OF PCMOAT SCORES OBTAINED

Program section	Condition	PCMOAT score			
		15	25	35	45
Punched cards	Constructed response	109	96	83	59
	Reading only	60	56	52	48
Card punches	Constructed response	156	143	130	116
	Reading only	69	64	60	55
Sorters	Constructed response	181	162	142	122
	Reading only	81	79	77	75
Reproducers	Constructed response	274	251	228	205
	Reading only	128	120	112	105

TABLE 4
ANALYSES OF COVARIANCE OF TIME IN MINUTES TO COMPLETE PROGRAMED TEXTS

Source	Punched cards		Card punches		Sorters		Reproducers	
	<i>df</i>	<i>F</i>	<i>df</i>	<i>F</i>	<i>df</i>	<i>F</i>	<i>df</i>	<i>F</i>
Constructed response versus reading	1	116.45**	1	548.46**	1	256.47**	1	445.80**
Error (<i>MS</i>)	(100)	(219.96)	(164)	(383.23)	(164)	(768.29)	(139)	(1053.03)
Heterogeneity of regression	1	7.56**	1	4.85*	1	13.36**	1	4.59*
Within groups regression (<i>MS</i>)	99	(206.42)	163	(374.45)	163	(714.45)	138	(1026.53)

* *p* = .05.
** *p* = .01.

Table 3 summarizes the times for subjects with particular PCMOAT scores to complete the programed materials. Again the values were estimated by least squares linear regression—in this table, time in minutes to complete a program section on PCMOAT score. The differences in time to complete the two editions of each program section were in every case very large, and the differences were larger for subjects with low PCMOAT scores than for subjects with high PCMOAT scores. Inspection of Table 3 also shows that the range of times to complete the programed materials was larger if a constructed response was required of the subjects than if they only read the materials.

The analyses of covariance of time in minutes to complete the program sections are summarized in Table 4. These analyses disclosed significant differences in mean time to complete the programed materials between the subjects that constructed a response and the subjects that only read the material for every program section. In addition to the differences in mean time, significant heterogeneity of regression was found between the two editions for all program sections.

A word of caution is in order for interpreting the heterogeneity of regression. When the dependent variable is a ratio scale, such as time, it is not uncommon to find the variance of a set of scores correlated with the mean. Under these circumstances, one would expect significant heterogeneity of regression whenever the treatments produced marked differences in group means. It appears though

that treatments in the present study did produce differences in slope beyond the expected differences due to the differences in means; when the slopes were adjusted for differences between treatment means the slopes for the groups constructing the response were still at least 30% greater than for the groups reading the materials only.

DISCUSSION

Quiz Scores. Finding no difference in mean performance on a criterion quiz between the subjects constructing a response and the subjects reading comparable material confirmed the data of most investigators (Campbell, 1961; Coulson & Silberman, 1961; Goldbeck, Campbell, & Llewellyn, 1960; Roe, Massey, Weltman, & Leeds, 1960). The results of the present study extend the work of previous investigators in that they were based on sizable program sections differing in internal structure, and the number of students used was large enough to get reasonable sensitive tests. "Acceptance" of the hypothesis of no differences between treatments is an unpalatable situation; one has not proved, nor can it be proved, that a null hypothesis is "true." Small scale, insensitive experiments make acceptance of a null hypothesis the rule. Thus when a null hypothesis is not disproved, some further examination of the data is desirable. One procedure is to determine the precision of the experiment by estimating the confidence interval for the difference between the treatment means. In the present study the standard error of the difference between

the means was approximately 2.10 percentage units, so reasonable precision was obtained. Considering the size of this interval with the nonsignificant differences on the criterion quiz, it appears safe to say that the mean difference in quiz scores between subjects using the two editions in this study was negligible for practical purposes. The interpretation of the slope of the regression lines is not as simple. There was significant heterogeneity of regression for only the program section dealing with the operation of the sorters; however, the same tendency was evident with the other three program sections. While the possibility exists that the significant F for the sorter program represents a Type I error, this appears unlikely as the same tendency was noted in the other program sections of equivalent length. The differences between the slopes of the regression lines for the two editions of the sorter program and reproducer program were not likely events; differences as large as those found would appear only with probabilities of .11 and .08, respectively. Unfortunately, the reason for the occurrence of a significant difference between the editions of the sorter program is not obvious; it was neither the easiest nor most difficult, the shortest nor the longest, and in terms of internal organization it was very similar to the reproducer program.

Time Scores. The time scores show large mean differences and heterogeneity of regres-

sion between subjects constructing a response and subjects only reading the materials. The practical importance of the smaller range in times for the program sections that require reading only rather than constructing a response cannot be overemphasized if programmed materials are to be used in classrooms rather than for home study.

REFERENCES

- CAMPBELL, V. N. Adjusting self-instructional programs to individual differences; studies of cueing, responding, and bypassing. *Amer. Inst. Res. tech. Rep.*, 1961, No. AIR-C41-7/61-SR.
- COULSON, J. E., SILBERMAN, H. F. A series of six studies in automated teaching. Paper read at National Education Association, Chicago, 1961. (Mimeo)
- GOLDBECK, R. A. The effect of response mode and learning material difficulty on automated instruction. *Amer. Inst. Res. tech. Rep.*, 1960, No. AIR-328-60-IR-124.
- GOLDBECK, R. A., CAMPBELL, V. N., & LLEWELLYN, J. E. Further experimental evidence on response modes in automated instruction. *Amer. Inst. Res. tech. Rep.*, 1960, No. AIR-328-60-R-132.
- HOLLAND, J. G. Design and use of a teaching machine. Paper read at American Psychological Association, Chicago, 1960. (Mimeo)
- ROE, A., MASSEY, M., WELTMAN, G., & LEEDS, D. Automated teaching methods using linear programs. *U. Calif. Los Angeles Rep.*, 1960, 60-105.
- SILVERMAN, R. E., & ALTER, M. Note on the response in teaching machine programs. *Psychol. Rep.*, 1960, 7, 496.
- SKINNER, B. F. Teaching machines. *Science*, 1958, 128, 969-977.

(Received October 22, 1962)

PREDICTING CREDIT RISK WITH A NUMERICAL SCORING SYSTEM

JAMES H. MYERS

School of Business Administration, University of Southern California

A "weighted application blank" approach was used to develop a numerical scoring system for predicting credit payment potential from personal history and financial obligation information. 17 biodata items from the credit application form were examined by chi square analysis to determine which items were discriminating between: (a) those previously accepted vs. those not accepted, and (b) among those accepted, those who paid vs. those who did not pay. Results showed good prediction of accepted vs. rejected applicants ($r_{pb1} \approx .60$), moderate predictions of paid vs. delinquent among those who were accepted ($r_{pb1} \approx .30$).

The application of numerical weights or scores to biographical data (personal history) items has long been familiar in the selection of personnel for industry. More recently, this approach has come to be used also in the prediction of credit risk at the retail store level. This paper reports the results of such a study in the central credit office of a large department store chain in Los Angeles, California.

The earliest published study dealing with the development of a numerical rating system was that by Durand (1941) under the sponsorship of the National Bureau of Economic Research. Hundreds of both good and bad personal loan accounts were analyzed from the files of commercial banks, personal finance companies, industrial banking companies, automobile finance companies, and appliance finance companies. Durand developed several different weighting systems for more accurately determining payment potential of an individual applying for credit. Results showed good prediction of credit repayment in 20 commercial banks and in 9 industrial banking companies (in aggregates, not as individual organizations). It is not certain, however, that these results were checked out on new samples to determine possible "shrinkage."

A later study by Wolbers (1949) was done in one branch store of a nationwide department store chain. Developing scoring weights on one sample and applying them to a second sample, he showed that credit losses could be reduced approximately 7%, with negligible losses in the volume of good business. Myers

and Cordner (1957) studied credit accounts in one branch of a Los Angeles chain dealing in personal loans. Results showed that approximately 6% of the losses from this branch could be eliminated with no losses in the volume of good business, approximately 24% of losses could be reduced if 3% of good business volume could be sacrificed, and approximately 50% of losses then experienced could be eliminated at the cost of only 7% of good business volume.

Another study (McGrath, 1960), conducted for an automobile dealer, found that approximately 20% of credit losses could be eliminated at the cost of only 1% or so of good business. Other studies (unpublished) known to the author include two for savings and loan institutions in the Los Angeles area and one for a personal loan company other than the one mentioned above (Myers & Cordner, 1957).

METHOD

The primary purpose of the present study was to develop a scoring system which would discriminate between good (open) and bad (delinquent) credit accounts. However, it was necessary to consider also those applicants who had been rejected, since prior experience in studies of this general type (Myers & Errett, 1958) have indicated the importance of determining the amount and direction of preselection (i.e., screening out applicants by credit evaluators on a judgment basis).¹ Also, it was of

¹ For example, in the present study monthly income did not prove to distinguish between good and poor accounts among those to whom credit was granted. Thus, weights would not normally be developed for this item. Yet a study of rejected appli-

TABLE 1
TOTAL NUMBER OF ACCOUNTS AND SAMPLE SIZE FROM EACH CATEGORY STUDIED

Account type	Approximate total number	Sample size	
		Original	Cross- validation
1. Granted credit-paid	225,000	149	199
2. Granted credit-delinquent	4,500	166	203
3. Refused credit	— ^a	141	200

^a Not available.

interest to see how well a scoring system could discriminate between *accepted* and *rejected* applicants, to determine the extent to which a tool of this kind could be used by inexperienced evaluators to make acceptance decisions on the same basis as those made by more experienced evaluators.

Table 1 below shows the total number of accounts in the groups studied and the sample size selected from each group.

Samples were drawn for each group from master company files by a “systematic” sample (every *n*th case). For example, in drawing “original sample” cases from the delinquent files, approximately every twenty-seventh case was drawn, proceeding from the front (A) of the alphabetical file to the back (Z). Resulting samples could be considered random samples of company accounts in each group.

A chi square analysis was then made between groups to find personal history items related to the *acceptance* of the application by the credit supervisors based on judgment (Groups 1 and 2 versus Group 3) and to the payment potential of accepted accounts (Group 1 versus Group 2). Two types of weights were developed for each predictive item: (a) chi square weights, based on the numerical value of chi square for each category of each item;

cations showed that almost every applicant with a monthly income under \$300 had been refused credit. While it is not known to what extent these people would have paid, it seems prudent to go along with experienced credit judgment in this case by assigning a negative weight to incomes under \$300 for purposes of the scoring system.

(b) unit weights, giving all positive chi square weights of whatever size a value of 2, neutral weights a value of 1, and negative weights a value of 0 (see Table 2).

RESULTS

Table 3 shows the probability levels resulting from the chi square analysis of each personal history item. Weights were developed for each of the predictive items for both conditions (accept versus reject; paid versus delinquent), and cross-validation samples were scored using both chi square and unit weights. Tables 4 and 5 show the results from applying the two weighting systems to the new groups.

Accept versus Reject

It is apparent from Table 4 that a numerical-type rating system can be used with substantial accuracy to predict decisions of experienced credit evaluators in accepting or rejecting applications for credit. Point-biserial correlation coefficients of .57 and .61 are evidence of this. Such coefficients, along with their corresponding *t* ratios, are significant at well beyond the .001 level. A glance at the distribution of scores in Table 4 shows

TABLE 2
EXAMPLE OF CHI SQUARE ANALYSIS OF ITEM, INCLUDING WEIGHTS FOR EACH CATEGORY

No. accounts in well-known stores	Open		Delinquent		χ^2 weight	Unit weight
	χ^2	No.	No.	χ^2		
0	8.8	28	75	7.9	—8	0
1	.1	58	60	.1	0	1
2 or more	7.7	63	31	6.1	7	2
Total	16.6	149	166	14.1		

TABLE 3
CHI SQUARE VALUES AND PROBABILITY LEVELS FOR EACH ITEM

Item	Accepted versus rejected			Paid versus delinquent		
	χ^2	<i>df</i>	<i>p</i>	χ^2	<i>df</i>	<i>p</i>
1. Sex applying	12.3	2	.001	— ^a	— ^a	— ^a
2. No. dependents	— ^a	— ^a	— ^a	11.1	4	.02
3. Marital status	49.0	4	.001	9.6	4	.05
4. Time at present address	— ^a	— ^a	— ^a	8.2	2	.02
5. Time at present and past address	— ^a	— ^a	— ^a	10.0	4	.05
6. Have telephone?	— ^a	— ^a	— ^a	21.6	1	.001
7. Rent or own home	29.6	2	.001	28.9	2	.001
8. Total accounts in other stores	41.3	5	.001	— ^a	— ^a	— ^a
9. No. accounts in well-known stores	47.0	2	.001	30.7	2	.001
10. No. finance company accounts	— ^a	— ^a	— ^a	8.6	2	.02
11. Total amount of debt	— ^a	— ^a	— ^a	— ^a	— ^a	— ^a
12. Occupation	11.8	3	.01	10.7	3	.02
13. Total monthly income	86.3	5	.001	— ^a	— ^a	— ^a
14. Bank account	59.3	3	.001	— ^a	— ^a	— ^a
15. Total amount of debt/Total monthly income	— ^a	— ^a	— ^a	6.0	2	.05
16. Personal reference given?	12.3	1	.001	— ^a	— ^a	— ^a
17. Both work?	— ^a	— ^a	— ^a	— ^a	— ^a	— ^a

^a Not significant; *p* > .05.

TABLE 4
SCORE DISTRIBUTIONS FOR ACCEPTED VERSUS REJECTED APPLICANTS

Unit weights			Chi square weights		
Score	No. accepted	No. rejected	Score	No. accepted	No. rejected
14	11	1	91 & up	29	3
13	32	3	85–90	38	6
12	36	3	79–84	40	1
11	37	12	73–78	44	17
10	33	22	67–72	18	21
9	16	14	61–66	12	24
8	14	21	55–60	9	29
7	13	30	49–54	7	28
6	3	28	43–48	1	25
5	2	18	37–42	1	28
4	1	23	31–36	0	11
3	1	17	30 & below	1	7
2	0	6			
1	1	2			
Total	200	200		200	200
Mean	10.61	6.80		78.24	55.07
σ	2.26	2.73		11.93	15.06
<i>t</i> ratio	15.20			17.04	
<i>r</i> _{pbi}	.57			.61	
<i>p</i>	<.001			<.001	

TABLE 5
SCORE DISTRIBUTIONS OF PAID VERSUS DELINQUENT ACCOUNTS

Unit weights			Chi square weights		
Score	No. paid	No. delinquent	Score	No. paid	No. delinquent
16	3	0	35 & up	5	0
15	13	2	33-34	40	10
14	23	8	31-32	25	15
13	26	9	29-30	10	5
12	30	17	27-28	13	3
11	38	21	25-26	43	42
10	24	26	23-24	24	32
9	16	38	21-22	11	27
8	15	32	19-20	4	8
7	6	26	17-18	20	40
6	4	13	15-16	2	6
5	1	7	13-14	1	13
4	0	3	12 & below	1	2
3	0	1			
Total	199	203		199	203
Mean	11.34	9.17		26.38	22.10
σ	2.30	2.38		5.60	5.36
t ratio		9.30			7.82
r_{pbi}		.31			.26
p		<.001			<.001

only a very small amount of overlap at the upper and lower extremes of score distributions for accepted and rejected applications. This is true for both unit weights and weights based on chi square values. For example, by using chi square weights, approximately 95% of the rejected cases fall below the median score (79.5) of the accepted cases; a very similar result emerges using unit weights.

The results from this phase would seem to offer encouragement to the very largest retail credit installations by showing that numerical rating systems can enable a less experienced credit evaluator to make decisions on a large percentage of incoming applications with a substantial degree of accuracy, or at least a substantial degree of agreement with decisions of more experienced evaluators.

It is interesting to note from Table 3 that evaluators were not selecting on the basis of some items which actually did discriminate between good and delinquent accounts. For example, applicants with telephones were found to be much more likely to pay

($\chi^2 = 21.6$) than those without telephones, yet evaluators were not using this item in their subjective evaluation of credit applications.

Paid versus Nonpaid

The ability of the numerical rating system to discriminate between paid and delinquent accounts is shown in Table 5. It can be seen that prediction is substantially less ($r_{pbi} = .31, .26$) than in the case of accepted versus rejected applications. This is not too surprising when it is remembered that paid versus nonpaid discrimination by a numerical system is in *addition* to that already effected by experienced credit evaluators. It is a measure of the *improvement* over present methods offered by such a system.

The overlap here is somewhat greater than that for acceptance versus rejection. Using chi square weights, approximately 82% of the delinquent group fall below the median score (26.2) of the paid accounts; a very similar result emerges using unit weights.

By eliminating all applicants scoring 6 or below (unit weight scale in Table 5), approximately 24% of potentially delinquent accounts could be eliminated at a cost of only 5% of good business. With delinquent accounts running at the rate of about 2% of all accounts, such a cutoff would result in reducing the total volume of new accounts by only slightly more than 5%.

Chi Square versus Unit Weights

A side issue of lesser importance in this study involved a comparison of the prediction effectiveness of numerical weights based upon raw chi square values and those based upon unit weights. Some idea of the comparative sizes of these weights can be gained by the fact that raw chi square weights ranged from -30 to 15 (a 45-point spread), while unit weights always range from 0 to 2. With equal predictive effectiveness, the advantages of working with unit weights in the operating situation are obvious.

Looking again at Tables 4 and 5, it can be seen that while unit weights are slightly *less* effective in predicting accepted versus rejected accounts, they were slightly *more* effective in predicting paid versus nonpaid accounts. In the balance, then, there is no clear evidence of the superiority of the more cumbersome chi square weights over unit weights, so that the latter should certainly be preferable for the operating situation. A

study cited previously (McGrath, 1960) also compared the effectiveness of chi square and unit weights with similar results.

DISCUSSION

As automation becomes more available to credit management, studies of the type described here will be simpler to perform. Also, the necessary maintenance or upkeep on established systems can be accomplished almost routinely through predetermined computer programs.

It will be many years, if ever, before systems of the type developed in this study will be so effective as to replace the experienced credit evaluator. They may, however, be a valuable operating and training tool for the large-scale credit operation in the near future.

REFERENCES

- DURAND, D. *Risk elements in consumer installment financing*. (Study No. 8) Washington, D. C.: National Bureau of Economic Research, 1941.
- MCGRATH, J. J. Improving credit evaluation with a weighted application blank. *J. appl. Psychol.*, 1960, **44**, 325-328.
- MYERS, J. H., & CORDNER, W. Increase credit operation profits. *Credit World*, 1957(Feb.), 12-13.
- MYERS, J. H., & ERRETT, W. The problem of preselection in weighted application blank studies. *J. appl. Psychol.*, 1958, **43**, 94-95.
- WOLBERS, H. L. The use of the biographical data blank in predicting good and potentially poor credit risks. Unpublished master's thesis, University of Southern California, 1949.

(Received October 29, 1962)

KENNETH E. CLARK, Editor
University of Rochester

Table of Contents

An Approach to an Objective Criterion for Research Managers: Lloyd H. Lamouria and Thomas W. Harrell.....	353
Color Coding and Visual Separability in Information Displays: Sidney L. Smith.....	358
Use of an All Possible Combination Solution of Certain Multiple Regression Problems: Joseph M. Madden and Robert A. Bottenberg.....	365
Effects of Noise Level and Difficulty of Task in Performing Division: James F. Park, Jr., and M. Carr Payne, Jr.....	367
Work Satisfaction and Scores on a Picture Interest Inventory: Harold Geist.....	369
Influence of Multiple-Choice Answer Form Design on Answer-Marking Performance: Irwin Miller and Frank J. Minor.....	374
Influence of Simultaneous Variation in Size of Type, Width of Line, and Leading for Newspaper Type: Miles A. Tinker.....	380
Payment History as a Predictor of Credit Risk: Howard W. Hassler, James H. Myers, and Maurice Seldin.....	383
Job Attitudes in Management: IV. Perceived Deficiencies in Need Fulfillment as a Function of Size of Company: Lyman W. Porter.....	386
Compensatory Tracking with Differentiating and Integrating Control Systems: E. C. Poulton	398
Evaluation of Typewriting Proficiency Training: Preliminary Test Development: Leonard J. West and D. J. Bolanovich.....	403
Effects of Heightened Motivation on the Detection of Deception: Lawrence A. Gustafson and Martin T. Orne.....	408
Doubling the Rate of Signal Presentation in a Vigilance Task during Sleep Deprivation: D. W. J. Corcoran.....	412
A Comparison of Three Approaches to Criterion Measurement: Forest O. Bell, Alvin L. Hoff, and Kenneth B. Hoyt.....	416

This is the last issue of Volume 47.
Volume Title Page and Contents appear herein.

American Psychological Association

(averaging 19 men each) will be labeled Departments A, B, C, and D.

Procedure

The unit director was asked to list the company objectives and to weight them in relation to each other. He was then asked to consider "Which objective is most important?" and "How much more important?" A procedure described by Churchman, Ackoff, and Arnoff (1957) was used in weighting objectives and yielded the following:

Profit	.5
Diversification	.2
Growth	.2
Welfare	.1

Next, those departmental activities contributing materially to the attainment of one or more of the objectives were listed for each of the four departments. Tables 1, 2, 3, and 4 list varied activities for the different departments; but many activities were typical for all departments, such as processes developed, publications, and papers presented.

The departmental activities were then weighted for their relative contribution to company objectives. For example, in Department A, Patent Disclosures contributed most to the profit objective; whereas

for Department D, Processes was rated most important for the same profit objective. A separate weighting of activities was then made, related to each of the four company objectives; for example, while Patent Disclosures contributed most to the profit motive in Department A, Devices was more important to the diversification objective in the same department.

The next step translated the data to standard units; that is, multiplying each of the various activity units, such as "dollars spent" and "papers processed," by a *common factor* (normalizing factor) which would permit the summation of nonsimilar units in a single, mathematical equation. This *common factor* (or coefficient) was a ratio of actual contributions divided by a predetermined ideal standard for all activities with the exception of proposals.² Actual contribution data were obtained

² Proposals are unique in this study because they represent an investment in outlining a proposed contract which may or may not result in a contract award. The management under study was using for its actual contribution the average cost of proposals over the average value of the contract received as an indication of effectiveness. Consequently, a standard ratio was obtained from the average yield of prior proposals.

TABLE 1
MODEL FOR EVALUATION

Department A activities	Objectives and weights			
	Profit, .5	Diversification, .2	Growth, .2	Welfare, .1
Papers	$.0 \times \frac{20^a}{15}$	$.0 \times 1.33^b$	$.1 \times 1.33^b$	$.2 \times 1.33^b$
Publications	$.0 \times \frac{8}{6}$	$.0 \times 1.33$	$.0 \times 1.33$	$.3 \times 1.33$
Patent Disclosures	$.4 \times \frac{11}{15}$	$.3 \times .733$	$.2 \times .733$	$.5 \times .733$
Devices	$.3 \times \frac{\$140,000}{85,750}$	$.4 \times 1.63$	$.4 \times 1.63$	$.0 \times 1.63$
Processes	$.1 \times \frac{\$ 20,000}{100,000}$	$.3 \times .200$	$.3 \times .200$	$.0 \times .200$
Proposals	$.2 \times \frac{.263^c}{.361}$	$.0 \times .728$	$.0 \times .728$	$.0 \times .728$
Σ Contribution to each objective	.474	.186	.198	.080
	Total contribution			.938

Note.—Slide rule calculations.

^a All figures represent $\text{Activity Weight} \times \frac{\text{Actual Contributions}}{\text{Standard Contributions}}$.

^b The decimal equivalent of the normalizer factor in the profit column. This is a constant for each activity and is repeated in decimal form to facilitate reading.

^c $\text{Activity Weight} \times \frac{\frac{\text{Department Cost to Submit Proposals}}{\text{Value of Contracts Received}}}{\frac{\text{Cost to Submit Proposals from } N \text{ Departments}}{\text{Value of Contracts Received by } N \text{ Departments}}}$

TABLE 2
MODEL FOR EVALUATION

Department B activities	Objectives and weights			
	Profit, .5	Diversification, .2	Growth, .2	Welfare, .1
Papers	$.0 \times \frac{1^a}{15}$	$.0 \times .667$	$.1 \times .667$	$.3 \times .667$
Publications	$.0 \times \frac{0}{5}$	$.0 \times 0$	$.1 \times 0$	$.3 \times 0$
Patents	$.1 \times \frac{2}{10}$	$.0 \times .2$	$.2 \times .2$	$.3 \times .2$
Processes	$.5 \times \frac{\$200,000}{400,000}$	$.5 \times .5$	$.4 \times .5$	$.1 \times .5$
Proposals	$.1 \times .36^b$	$.1 \times \text{—}^b$	$.1 \times \text{—}^b$	$.0 \times \text{—}^b$
Materials Application	$.3 \times \frac{\$ 50,000}{167,000}$	$.2 \times .299$	$.1 \times .299$	$.0 \times .299$
Σ Contribution to each objective	.180	.062	.067	.031
	Total contribution			.340

Note.—Slide rule calculations.
^a All figures represent Activity Weight $\times \frac{\text{Actual Contributions}}{\text{Standard Contributions}}$.
^b Actual contribution not available because the time interval since origin had been too brief.

TABLE 3
MODEL FOR EVALUATION

Department C activities	Objectives and weights			
	Profit, .5	Diversification, .2	Growth, .2	Welfare, .1
Papers	$.0 \times \frac{10^a}{20}$	$.0 \times .50$	$.1 \times .50$	$.2 \times .50$
Publications	$.0 \times \frac{11}{8}$	$.0 \times 1.375$	$.1 \times 1.375$	$.3 \times 1.375$
Patent Disclosures	$.2 \times \frac{0}{10}$	$.2 \times 0$	$.1 \times 0$	$.5 \times 0$
Processes	$.5 \times \frac{\$155,000}{282,000}$	$.2 \times .55$	$.3 \times .550$	$.0 \times .550$
Proposals	$.0 \times \frac{.032}{.36}$	$.2 \times .839$	$.1 \times .839$	$.0 \times .839$
Materials Application	$.3 \times \frac{\$ 40,000}{142,900}$	$.4 \times .282$	$.3 \times .282$	$.0 \times .282$
Σ Contribution to each objective	.180	.078	.104	.051
	Total contribution			.413

Note.—Slide rule calculations.
^a All figures represent Activity Weight $\times \frac{\text{Actual Contributions}}{\text{Standard Contributions}}$.

TABLE 4
MODEL FOR EVALUATION

Department D activities	Objectives and weights			
	Profit, .5	Diversification, .2	Growth, .2	Welfare, .1
Papers	$.1 \times \frac{38^a}{50}$	$.0 \times .761$	$.1 \times .761$	$.0 \times .761$
Publications	$.1 \times \frac{1}{5}$	$.0 \times .200$	$.1 \times .200$	$.5 \times .200$
Patent Disclosures	$.0 \times \frac{14}{15}$	$.3 \times .933$	$.2 \times .933$	$.5 \times .933$
Processes	$.4 \times \frac{\$40,000}{46,200}$	$.3 \times .866$	$.3 \times .866$	$.0 \times .866$
Devices	$.3 \times \frac{\$33,600}{36,000}$	$.4 \times .933$	$.3 \times .933$	$.0 \times .933$
Tests	$.1 \times \frac{\$768,000}{768,000}$	$.0 \times 1.000$	$.0 \times 1.000$	$.0 \times 1.000$
Σ Contribution to each objective	.410	.183	.165	.057
	Total contribution			.815

Note.—Slide rule calculations.
* All figures represent Activity Weight $\times \frac{\text{Actual Contributions}}{\text{Standard Contributions}}$

from historical records. The ideal standard was based on the subjective judgment of the department manager and represents a theoretical goal deemed both desirable and feasible. Whenever the actual contributions equal the ideal (or expected), the normalizer factor equals unity. And, since the weights assigned to each activity sum to 1 (the same is true of the objective weights), final OR scores of unity indicate that the expected goal has been achieved whereas values of less than 1 may indicate need for improvement (see discussion for elaboration).

The final step was the combination of all factors to produce a single-value index.

Example: In Table 1 for Department A, the value of activities contributing to profits is:

Contribution to Profit = [Weight of Profit Objective]

$$\times \left[\Sigma \text{Activity Weight} \times \frac{\text{Actual Contributions}}{\text{Standard Contributions}} \right]$$

$$= .5 \left[\begin{array}{ccc} \left(.0 \times \frac{20}{15} \right) & + & \left(.0 \times \frac{8}{6} \right) + \left(.4 \times \frac{11}{15} \right) \\ \text{Papers} & & \text{Publications} \quad \text{Patents} \end{array} \right.$$

$$+ \left(.3 \times \frac{\$140,000}{85,750} \right) + \left(.1 \times \frac{\$20,000}{100,000} \right)$$

$$\begin{array}{cc} \text{Devices} & \text{Processes} \end{array}$$

$$+ \left[.2 \times \frac{\$14,623}{\frac{55,600}{29,467}} \right] \left[\frac{81,787}{81,787} \right] = .474.$$

TABLE 5
SUMMARY OF MODEL

Department manager	Indices of contribution to:				Single index
	Profit	Diversification	Growth	Welfare	
A	.474	.186	.198	.080	.938
D	.410	.183	.165	.057	.815
C	.180	.078	.104	.051	.413
B	.180	.062	.067	.031	.340

TABLE 6

COMPARISON OF OR RATING AND RANK
ORDER BY CLINICAL RATING

Department Manager	Rank order by:	
	OR rating	Clinical rating
A	1	3
D	2	1
C	3	4
B	4	2

This operation is repeated for each of the other three objectives and the figures summed:

Total contributions = contribution to Profit + Diversification + Growth + Welfare = .474 + .216 + .228 + .103 = 1.02.

RESULTS

The input data and results for each of the four departments are shown in Tables 1, 2, 3, and 4. A summary and comparison of results for the four departments are shown in Table 5. Finally, a comparison of OR analysis and clinical rating is shown in Table 6.

The conclusions drawn from the OR scores are as follows:

1. The contributions surrounding the manager of Department A made possible a score about equal to the expected or ideal level of performance, that is, 1.00.

2. In contrast, the performance of the other three managers was below expectation; Department B contributed only about $\frac{1}{3}$ of the desired level of performance.

DISCUSSION AND CONCLUSIONS

OR makes possible an empirical evaluation of a manager's contribution. Hard-to-quantify data are molded into a form yielding consistent measures, reducing clinical judgments that normally accompany management evaluation—thus this technique (model building) provides management with a tool of exceptional analytical qualities and permits convenient quantification.

The director of the four departments was asked to place the departments in an ordinal

ranking. This was done according to his clinical evaluation of their performances relative to attainment of goals which he hoped they would achieve. A comparison of the OR ratings and the director's ordering of the four departments is shown in Table 6. The authors are convinced that the lack of correspondence shown between the global ratings and the ratings obtained from the empirical objective approach of OR demonstrates the low validity of the clinical, subjective approach to management evaluation.

Because the OR model is merely a reflection of centralized control, it readily lends itself to detailed study and isolation of weak areas in management contribution. But, by the same token, it is important that the model be continually reviewed to insure that the programing (selection of standard ratios) is correct. The final scores are meaningful only if the standards are adequate.

The model evaluates the manager in such an all-inclusive manner that frequently factors beyond his control are reflected in the index. On the other hand, perhaps a manager should be responsible for his group's actions. An able manager who receives an unfavorable index owing to factors beyond his control will be quick to use this model in his own defense. He may now demonstrate to his superiors that some of the executives' acts are at cross purposes with the desired model and prevent attainment of company objectives.

REFERENCES

- CATTELL, R. B. New concepts for measuring leadership in terms of group syntality. In D. Cartwright & Alvin Zander (Eds.), *Group dynamics*. Evanston, Ill.: Row, Peterson, 1953. Pp. 14-28.
- CHURCHMAN, C., ACKOFF, L., & ARNOFF, E. *Introduction to operations research*. New York: Wiley, 1957. Pp. 150-152.
- HEIDER, F. Social perception and phenomenal causality. In R. Tagiuri & L. Petrullo (Eds.), *Person perception and interpersonal behavior*. Stanford: Stanford Univer. Press, 1958. Pp. 1-21.
- HERMAN, C. C., & MAGEE, J. F. Operations research for management. *Harv. Bus. Rev.*, 1953, 31, 101.
- ROBERTS, F. B. Operational research, managements need. *Engineering*, 1961, 192, 652-653.

(Received October 29, 1962)

COLOR CODING AND VISUAL SEPARABILITY IN INFORMATION DISPLAYS¹

SIDNEY L. SMITH

MITRE Corporation, Bedford Massachusetts

12 experimental Ss performed both visual search and class counting tasks, viewing displays containing 20, 60, or 100 items. Each item consisted of a vector, letter, and 3-digit number grouped together, and was presented as white-on-black in some displays, or in 1 of 5 colors. The color code was redundant with the 5 class-designator letters that were used. Average search and counting time, and counting errors, increased with increasing display density (number of items). None of these measures varied significantly among the 5 different target classes (colors). Addition of the redundant color code resulted in an average time reduction of 65% in the visual search task and 69% in the counting task, with a reduction of 76% in counting errors.

Many factors must be considered in assessing the value of color coding in information displays: the legibility of colored symbols under differing viewing conditions, the number of colors that can be reliably identified on the basis of absolute discrimination, and the interpretability of color codes in different display formats. The results of experimental studies dealing with these questions have been summarized in the recent review article by Jones (1962).

Among the most basic experimental studies with implications for display design are two that estimate the degree of visual separability among displayed data classes provided by color coding in the context of a visual search task—the first by Green and Anderson (1956) and the later replication by Smith (1962). These studies confirmed that the visual separability provided by a code of as

many as four (or five) colors is almost perfect. That is to say, observers were able to scan displayed items of one color class almost as quickly in the presence of different-colored items as when they were presented alone.

These studies are basic in the sense that visual search underlies much of the use of information displays in practical situations. However, the simple search tasks used in these experiments are far from approaching in complexity the many ways in which displays can be used in the actual context of man-machine systems. An extension of this work is necessary to determine to what degree the visual search data can be used to predict operator performance in more complicated tasks involving the use of displays.

This present study represents a direct extension of the visual search model to a task involving class counting, where an observer must note *all* the displayed items of a particular class rather than find just one of them. Counting is itself also a relatively simple task compared with many actual display situations, but it does represent a next step in the direction of increased task complexity.

The intent, of course, was not just to confirm the value of color coding for a counting task. That color coding permits faster counting might be predicted with considerable confidence on the basis of the search task data already at hand. The purpose was rather to obtain direct estimates of the *extent* of improvement in performance that can be at-

¹ The research reported in this article was supported by the Air Force Electronic Systems Division, Air Force Systems Command, under Contract AF-19(628)2390. A more detailed account of this research was published as a MITRE technical series report, MTS-10: "Display Color Coding for Visual Separability," August 1963.

The author wishes to acknowledge the help of Mari R. Jones, who assisted in the preparation of experimental displays; of David E. Moore, who was largely responsible for the data collection; and of Barbara B. Farquhar, who assisted in the data analysis.

² The words "random" or "random selection," as used in this report, are intended in every instance to indicate an unbiased selection from among equiprobable alternatives, with any qualifying restrictions indicated as such.

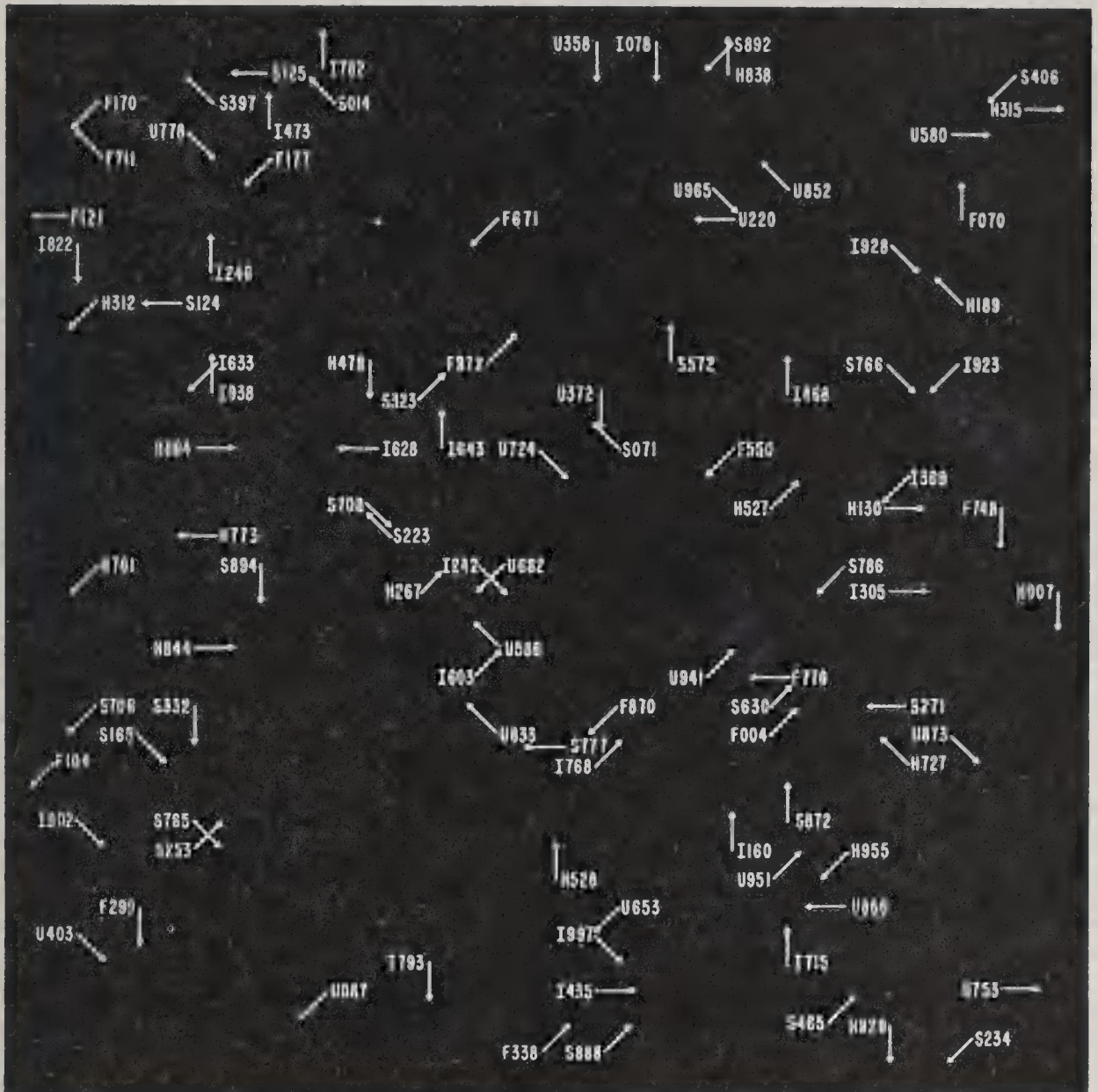


FIG. 1. Sample display of 100 items with, of course, no color coding.

tributed to color coding, and in particular to determine the degree to which this improvement in counting is consistent with the data obtained in visual search tasks.

PROCEDURE

Twelve men and women with normal color vision served as experimental subjects. In the course of the study, each subject worked individually, viewing in sequence a total of 180 displays in two experimental sessions. The displays were prepared as 2×2 inch slides and presented by rear projection on a viewing screen to produce a 32-inch square display field. The subjects observed the screen from a distance of 4 to 5 feet, depending

upon momentary sitting posture, with the center of the display field approximately at eye level. The ambient illumination was moderately high, about 2 foot-candles.

Each display presented a number of items, described to the subjects as "tracks," shown as white (or colored) on a dark background. A displayed item consisted of a group of symbols including a letter designating the track class (one of five: F, I, S, H, or U), followed by a 3-digit number and an adjacent vector (arrow) pointing in one of eight directions. In each display the items were placed randomly² in the display field by selection from among 324 possible positions representing the intersections of 18 columns by 36 rows. Displays were prepared containing 20, 60, and 100 items to permit measurement over a range of display

densities. A sample display of 100-item density is presented in Figure 1. The character height of the alpha-numeric symbols as projected on the screen was 14 millimeters and the stroke width 3 millimeters.

In preparing the displays, a class-designator letter was selected at random for each item with the provision that there be at least one item of each class on every display. The numerical designator was also chosen randomly for each item, as was the vector direction. The vectors had no functional significance in this study, but were simply added as "clutter" to increase apparent display complexity. The numerical designators had to be noted by the subjects in the search task, and the class-designator letters were important in both search and counting tasks, as will be described later.

In half of the displays used, there was no color coding; the items were white on a dark background. In the other displays, all the symbols in an item were displayed in a particular color with the color code chosen to be redundant with the class-designator letter: items containing an "F" were green, "I" white, "S" blue, "H" red, and "U" yellow. Displays were prepared in pairs, with and without the color code, so as to be identical in other respects. Nine of these pairs of displays were made for each of the three display density levels.

The apparent colors of the displayed items as projected are described below in terms of Munsell color specifications. Also listed are estimated brightness levels of the different colors as measured by a spot photometer.

Display color	Munsell notation	Approximate brightness (foot-lamberts)
Green	2.5 G 6/8	20
White	5 Y 8/4	100
Blue	10 BG 6/6	20
Red	7.5 R 4/12	10
Yellow	7.5 YR 6/10	60

In half of the experimental trials the subjects were asked to scan the display to find a particular target item and the performance measure taken was search time. Subjects were instructed to search as quickly as possible. A subject was told in advance the letter and first 2 digits of the target item and indicated when he had found the target by reading aloud the third digit. The particular display to be used and target item to be searched for were chosen by the experimenter randomly with the restriction that the letter-plus-2-digit indicator had to define a unique item on the display. During the course of the experiment each subject had to search, in different trials, for items of each of the five possible target classes, on displays of each of the three density levels, with and without color coding. Moreover, for each of these conditions three replicate search measures were taken, using three different display slides randomly chosen for each subject.

In the other half of the experimental trials, randomly interspersed with the search trials, the subjects were asked in advance not to find a particular item but rather to count *all* the items of a designated target class and report the total number displayed. Subjects were instructed to count as quickly and accurately as possible. Performance measures taken were counting time and error if any. The subjects were given no information during the experiment as to the accuracy of their counts. Again, in various trials during the experiment each subject had to count items of each of the five classes, in displays at each of three density levels, with and without color coding, with three replicates (using different displays) under each condition. A random order of display presentation and task sequencing was chosen separately for each subject.

In summary, the experimental design represents a factorial combination of 12 subjects, by three display densities, by two types of display (with or without color coding), by five target classes, by two types of task (search or count), with three replicates under each condition. This provided a total of 1,080 time measures for the search trials and both time and error scores from the 1,080 counting trials.

RESULTS

The results of this study can be summarized as follows. Both time and error scores increased with increasing display density. Counting took longer than searching. Color coding markedly reduced both search and counting times as well as counting errors.

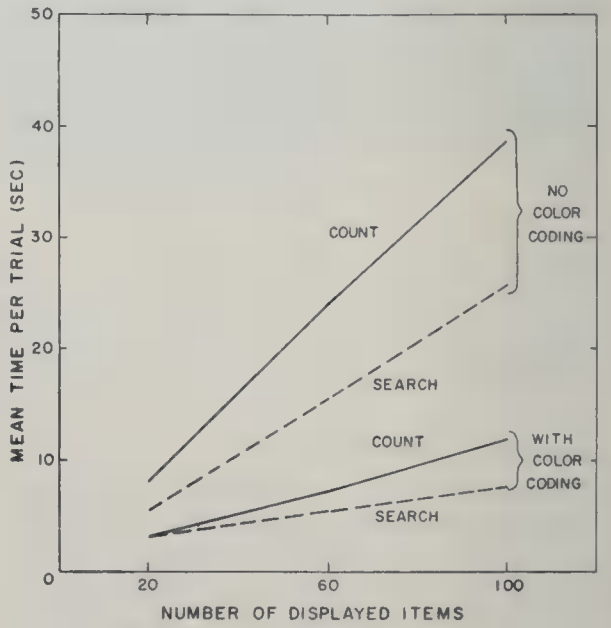


FIG. 2. Average counting and search times as a function of display density with and without color coding.

There were no significant differences among the several target classes.

The increase in search and counting time for displays of increasing item density is illustrated in Figure 2. This increase appears to be linear over the display range used, which is consistent with earlier results (e.g., Smith, 1962). Also illustrated in Figure 2 is the sizable difference in time required for the performance of the two different experimental tasks. This difference is in the expected direction: counting, when the subjects had to scan all the displayed items, took longer than the search task. The influence of color coding is clear—it made both tasks easier.

Table 1 summarizes the results of an analysis of variance performed upon time scores, where a score was the average of the three replicates for an individual subject under a particular combination of experimental conditions: working on either the search or counting task, for each of the five target classes, in displays at each of the three density levels, with and without color coding. The interpretation of the analysis is a conservative one, that advocated by Edwards (1950) for studies involving repeated measurements from the same subjects. No con-

clusions are drawn concerning possible individual differences among the subjects, or interactions between subjects and the various experimental conditions.

This analysis confirms the statistical reliability of time differences between the two tasks, those related to display density, and those attributable to the presence or absence of the redundant color code. Also confirmed as statistically significant are interaction effects between display density and task, between display density and coding, and between task and coding. Of lesser magnitude but perhaps also reliable ($p < .01$) is the three-way interaction among these same experimental variables, task, display density, and coding. No significant differences were confirmed among the various target classes (letter codes in the black-and-white displays, letter-plus-color codes in the colored displays), nor did this variable of target class interact significantly with any of the other experimental conditions.

Because of the correlation that typically occurs between the mean and variance of search time scores, several parallel variance analyses were also made of the time data. (The availability of automatic data processing equipment made this a relatively easy thing to do.) One of these alternative analyses was also based on average time scores, but used a logarithmic transformation of the scores. Two other analyses were based on the original scores themselves, taking the three replicates in each condition separately, one analysis with a log transform of these scores, and one without. All of these supplementary analyses confirmed the results of the one summarized here, with F ratios that were, if anything, somewhat higher.

The value of the display color coding in the counting task was not limited to its effect in reducing time required. There was also a sizable reduction in counting errors, illustrated in Figure 3. As compared with the black-and-white displays, the color coded displays resulted in an overall error reduction of 76% in the counting task. Figure 3 also shows an increase in counting errors with increasing item density, corresponding to the longer counting time required for these more dif-

TABLE 1
ANALYSIS OF VARIANCE FOR TIME SCORES

Source of variance	<i>df</i>	<i>MS</i>	<i>F</i>
Subjects (Ss)	11	355.6	—
Conditions	59	—	—
Task (T)	1	4,306.6	130*
Color Code (C)	1	31,332.8	946*
Display Density (D)	2	15,533.3	469*
Target Class (L)	4	19.3	.58
T × C	1	1,678.5	50.7*
T × D	2	808.6	24.4*
T × L	4	52.4	1.58
C × D	2	5,132.8	155*
C × L	4	61.6	1.86
D × L	8	39.1	1.18
T × C × D	2	156.8	4.73
T × C × L	4	36.8	1.11
T × D × L	8	54.0	1.63
C × D × L	8	52.2	1.58
T × C × D × L	8	28.1	.85
Ss × Conditions	649	33.1	—

* $p < .001$.

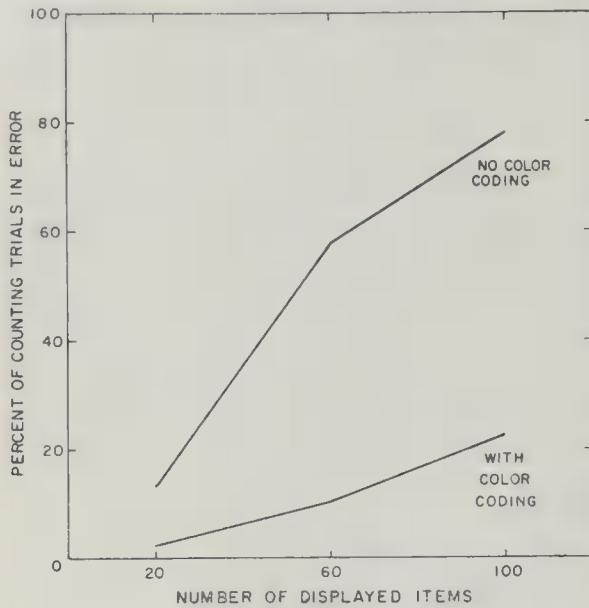


FIG. 3. Counting errors as a function of display density with and without color coding.

ficult displays. Most of the errors made, 88% overall, were those of omission, that is, underestimates rather than overestimates. It also happened that under conditions where errors were more frequent there was an increased likelihood of larger errors. However, an analysis based simply on error frequency, without regard to sign or size of error, is sufficient to confirm statistically reliable differences among the various conditions.

In general, the error frequency data tend to confirm the overall conclusions based on counting time. Chi square analysis of the relative frequency of errors versus correct counts confirmed that color coding reduced error frequency as compared with black-and-white displays ($\chi^2 = 182.4$; $df = 1$, $p < .001$) and that errors occurred more frequently as display density was increased ($\chi^2 = 157.0$; $df = 2$, $p < .001$). Differences attributable to the particular target class were much less marked. Although chi square analysis suggested there might be a possible effect here ($\chi^2 = 11.5$; $df = 4$, $p < .05$), most of this difference was contributed by data from the black-and-white displays and so cannot be attributed to the differences among the code colors used.

Examining briefly the question of individual differences in performance among the experimental subjects, a rank-order correla-

tion between the overall search times and counting times for each subject indicated a consistent relation ($\rho_s = .67$; $p < .05$). In terms of overall frequency of counting errors, chi square analysis confirmed that there were reliable differences among the 12 subjects ($\chi^2 = 40.6$; $df = 11$, $p < .001$). There was some indication of a consistent relation between the number of errors made by each subject for the black-and-white displays and his error frequency for the color coded displays ($\rho_s = .64$; $p < .05$). There was no discernible relation between overall counting time and errors ($\rho_s = .08$).

DISCUSSION

The marked dependence on display density of search time and counting time (and errors) certainly corresponds to our practical experience that "crowded" displays, those containing a great deal of data, are difficult to work from quickly and accurately. Indeed, this relation between ease of display interpretation and display density is so fundamental that the effects of other display variables—changes in format and/or of coding dimensions—may perhaps best be described in terms of their influence on this basic function. This point is illustrated in Figure 2, where it may be seen that the sizable decreases in search and counting time as a result of adding redundant color coding can be economically described as changes in the *slope* of the linear function relating time required to display density. This postulated relation is lent some support by the statistical analysis, which confirms not only the reliability of the overall differences attributable to color coding but also a significant interaction of this effect with display density.

This evidence of the effectiveness of color coding is not surprising under the circumstances of this particular experiment. In the black-and-white displays, the single letter which was the class designator was a relatively inconspicuous portion of each displayed item, whereas when the entire item was color coded its class was apparent at a glance, permitting more rapid display scanning. That is to say, color coding increased discriminability among items of the various classes. This improved visual separability has

the effect of making the spatial distribution (patterning, clustering, etc.) of the randomly placed items more perceptible and guides the eye during search and counting tasks. Presumably this effect of visual separation is limited by the discriminability of the colors used. If more and more different colors were added to a display, one would expect that there would be diminishing improvement followed by eventual degradation of visual separability based on this coding dimension.

The average search times in the present experiment are somewhat greater than those obtained in a comparable earlier study (Smith, 1962) which used five-color redundant coding with 3-digit number items. This may reflect some detrimental "clutter" effects of adding vectors and class identifying letters to the displayed items in the present study. However, this is speculative since there were other incidental differences between the two studies: different shaped numbers, slightly different hues, differences in subtended visual angle of the displayed items, and in ambient illumination. And, as it happened, most of the subjects were different in the two studies. The important thing to note is that the relative value of color coding in reducing search time was consistent in both studies in spite of these various procedural differences. The percent reduction in search time attributable to color coding in the two studies was:

Display density (number of items)	First study (Smith, 1962)	Present study
20	43%	45%
60	65	64
100	68	70

Another way of stating the same comparison is to note that in both studies the slope of the linear relation between average search time and display density was reduced by the same amount, about 75%, with the addition of redundant color coding.

The reduction in *counting* time attributable to color coding, for displays of 20, 60, and 100 items, was 63, 70, and 69%, respectively. The reduction in slope relating counting time to display density was about 75%. Thus the value of color coding in the counting task was comparable to its value in the visual

search situation. This finding suggests that the effect of color coding in actual display applications may eventually also prove to be predictable from such basic experimental measures.

Of course, there were obvious differences in performance time required by the search and counting tasks. Again, the effect can be economically described as a difference in the slope of the linear function relating time required to display density. Both the overall difference between tasks and the interaction of this effect with display density were confirmed as statistically significant.

It is interesting that in this study the average counting time, although greater than search time, was not twice as great—as one might predict on the basis of a hypothetically efficient search in which on the average only half of the possibly relevant items are scanned. In practice, the subjects were somewhat less efficient in their search procedures, occasionally overlooking the target item and having to scan repetitively. For the 100-item displays, average search time was approximately two thirds of the corresponding counting time, both for the color coded and black-and-white displays. The fact that counting was just as fast as search for the 20-item color coded displays may be a transient effect related to the possibility of subitizing (Cf. Kaufman, Lord, Reese, & Volkmann, 1949, p. 520) the number of displayed items when there are only a few of them and they are readily apparent.

In view of the conclusion reached above that color coding produces an equivalent reduction in time in search and counting tasks, it might appear somewhat anomalous to note that the statistical analysis confirms a significant interaction effect between task and color coding. The explanation lies in the fact that the statistics in this instance are those relating only to the absolute levels of performance, as measured by time scores; and it happened that both tasks were so much easier with the colored displays that the absolute difference between search and counting times was less under these circumstances than for the black-and-white displays. The interpretation given above, in terms of the *proportional* reduction in time scores, makes

the task equivalence of the effect of color coding much more apparent.

The failure to find significant differences among the various letter-color-classes is consistent with the conclusion already published (Smith, 1962) that the particular color of a displayed target does not, per se, necessarily influence search performance. However, through a fault in the experimental design the number of target items actually displayed in each class varied somewhat from subject to subject, and from one condition to another, depending upon the random choice among available display slides. This inadvertent confounding with display "sub-density" may have obscured small differences among the target classes. In other circumstances differences in legibility among different displayed colors have been confirmed, for example, under conditions of symbol overprinting (Smith, 1963). Thus one might well

choose to remain open-minded on the subject. Certainly the present data suggest that if differences in displayed target color are discovered for search and counting tasks, these differences will be relatively small in magnitude.

REFERENCES

- EDWARDS, A. L. *Experimental design in psychological research*. New York: Rinehart, 1950.
- GREEN, B. F., & ANDERSON, LOIS K. Color coding in a visual search task. *J. exp. Psychol.*, 1956, **51**, 19-24.
- JONES, MARI R. Color coding. *Hum. Factors*, 1962, **4**, 355-365.
- KAUFMAN, E. L., LORD, M. W., REESE, T. W., & VOLKMANN, J. The discrimination of visual number. *Amer. J. Psychol.*, 1949, **62**, 498-525.
- SMITH, S. L. Color coding and visual search. *J. exp. Psychol.*, 1962, **64**, 434-440.
- SMITH, S. L. Legibility of overprinted symbols in multicolored displays. *J. engng. Psychol.*, 1963, **2**, 82-96.

(Early publication received July 24, 1963)

USE OF AN ALL POSSIBLE COMBINATION SOLUTION OF CERTAIN MULTIPLE REGRESSION PROBLEMS¹

JOSEPH M. MADDEN AND ROBERT A. BOTTENBERG

Aerospace Medical Division, Lackland Air Force Base, Texas

The advent of electronic computing facilities has made the computation of an R^2 for all possible combinations of a set of predictors practicable. If there are several subsets of predictors with approximately equal predictive efficiency, the choice among subsets may be based upon such additional considerations as face validity, ease of measurement, and dimensionality. In the analysis reported in this paper, a choice could be made from among 6 subsets of predictors with practically no loss of predictive efficiency and from among 68 subsets if some loss were acceptable.

In the December 1955 issue of the *American Psychologist*, Ward (1955) discussed the Wherry-Doolittle method of computing multiple correlation. He suggested that an alternative analysis might be desirable in which the squared multiple correlations and beta weights are computed for the n predictors in all possible combinations of from 2 to n . Such computations, Ward indicates, would allow a better understanding of the interrelations of the predictor variables.

The computation of R^2 for all possible combinations of a set of predictors is quite a task for even a large electronic computer when n is very large, as Ward points out. However, for small values of n the task is quite practicable, even for small computers.

The computation of all possible combinations of a set of predictors is particularly pertinent in the establishment of a job evaluation plan. The initial procedure usually followed is to identify all the job requirement factors which appear to be present in the population of jobs to which the plan is to be applied. Some method is then used to identify the most efficient subset of this initial set. Lawshe and Maleski (1946) describe a typical study in which the Wherry-Doolittle solution was used. They concluded that 3 primary factors out of a full set of 11 accounted for practically all of the variance (96%) in total evaluation score.

The data from the Lawshe and Maleski study were used for an all possible combination solution in order to observe the kind of additional information which might be obtained. The results are in Table 1. The table should be read as follows: the entries under the distribution of MCC (multiple correlation coefficient) squared indicate the number of combinations in the interval from the indicated column to the next lower column. For instance, there are 172 combinations of four predictors which yield an R^2 of .90-.95. The first point of interest is that there are six subsets of three predictors which account for 95-96% of the total predictable variance. Since the best set of three predictors accounts for .9601 of the variance, there is little difference in the efficiency of the six subsets of three predictors. In addition, there are 68 subsets of three predictors which account for between 90% and 95% of the total variance. For these data, decisions concerning the selection of factors for the job evaluation plan can be based on a good deal more information than that available for Lawshe and Maleski on the basis of the Wherry-Doolittle solution before electronic computers were commonly available. Having a choice among several subsets of factors is important because some factors possess more face validity than others, some lend themselves more readily to scaling, some are simpler to use because of their unidimensionality, and so on.

The stability of the rank order of predictive efficiency of the various combinations

¹The work reported in this paper was sponsored by the 6570th Personnel Research Laboratory, Aerospace Medical Division, under Air Force Systems Command Project 7734 (02).

TABLE 1
ALL POSSIBLE COMBINATION SOLUTION FOR 11 PREDICTORS

Number of predictors	Distribution of MCC squared																				Best predictor set	Highest MCC ²
	.05	.10	.15	.20	.25	.30	.35	.40	.45	.50	.55	.60	.65	.70	.75	.80	.85	.90	.95	1.00		
1	1		1					1			1	1	3					2			2	.8649
2			1					2		1	1	3	2	8	5	11	12	9			2, 3	.9300
3								1			1	3	1	4	10	27	44	68	6		2, 3, 10	.9601
4													1		1	5	18	68	172	65	2, 3, 6, 10	.9747
5																1	3	41	205	212	2, 3, 5, 6, 10	.9812
6																		8	114	340	1-3, 5, 6, 10	.9858
7																			25	305	1-3, 5, 6, 10, 11	.9911
8																			1	164	1-6, 10, 11	.9933
9																				55	1-7, 10, 11	.9943
10																				11	1-7, 9-11	.9948
11																				1	1-11	.9951

in Table 1 is, of course, a matter for a cross-validation study, and no implication is intended concerning the likelihood that each combination of predictors would retain either its absolute level or rank order of predictive efficiency if a different sample from the same population of jobs were used. Further, it should be noted that data for any particular subset which entered the all possible combination solution may have had some degree of experimental dependence on predictors which the all possible combination results indicate can be dropped, and it would be desirable to determine whether such omission does in fact leave a desired subset which will achieve the stated level of predictive efficiency.

Another aspect of the all possible combination solution shown in Table 1 is that the question of criteria for decision making is clarified. For instance, all of the increases in R^2 shown in the column at the far right

are statistically significant. However, some of the increases are so small as to have little or no practical significance. It seems clear that the question of how many factors to use could be best answered by computing evaluation scores using factor sets of 3, 4, 5, and so on. When there ceases to be any difference (practical or statistical) between evaluation scores computed with the complete set and the scores obtained by use of the subset, it can then be said that the most efficient subset has been identified.

REFERENCES

LAWSHE, C. H., JR., & MALESKI, A. A. Studies in job evaluation: II. An analysis of point ratings for salary paid jobs in an industrial plant. *J. appl. Psychol.*, 1946, 30, 117-128.

WARD, J. H., JR. Use of electronic computers in psychological research. *Amer. Psychologist*, 1955, 10, 826-827.

(Received October 8, 1962)

EFFECTS OF NOISE LEVEL AND DIFFICULTY OF TASK IN PERFORMING DIVISION

JAMES F. PARK, JR., AND M. CARR PAYNE, JR.

Georgia Institute of Technology

Ss worked division problems for 20 min. in the presence of 98 db.-108 db. of noise while Ss of comparable mean ability in arithmetic worked the problems under conditions of room noise. Under each condition 1 group of Ss worked "easy" problems and another group worked "difficult" problems. Intense noise produced no effect on mean number of problems correctly solved. Variability of performance was significantly greater with easy problems under intense noise conditions than under room-noise conditions, although there was no difference with difficult problems. There was no evidence of a decrement in performance within the 20-min. session attributable to noise level.

Decremental effects on task performance have appeared when the subject worked in the presence of noise levels greater than 90 decibels (Broadbent, 1955, 1958b), particularly if the task was not intrinsically challenging (Jerison, 1959). Easier tasks were less affected than more difficult ones. These effects have been shown in mental multiplication (Viteles & Smith, 1946) and in subtraction (Broadbent, 1958a).

The present study was designed to extend the research on effects of noise upon performance on an unchallenging intellectual task by examining effects of difficulty of the problem as well as noise and also by examining variability of performance.

PROCEDURE

Forty male college students served as subjects. On the basis of a 5-minute pretest consisting of mathematical problems subjects were divided into four groups. No significant differences among the groups were shown by *t* ratios computed on mean performance on the pretest.

The experimental session lasted for 20 minutes. Two of the groups (E groups) worked "easy" division problems (defined as problems containing single-digit divisors) and the other two groups (D groups) worked more "difficult" division problems (defined as those containing two-digit divisors). At 5-minute intervals a signal was given. At this time, each subject marked his paper beneath the last completed problem.

One E group and one D group worked their problems at the same time at room noise level (50

decibels to 70 decibels).¹ The other E group and D group worked their problems at the same time in the presence of 98-decibel to 108-decibel noise produced by an air horn² (three chimes—277 cycles per second, 329 cycles per second, and 440 cycles per second) placed 6 feet above floor level in the front of a modern classroom (23 × 25.5 × 10 feet).

RESULTS

The number of problems solved correctly served as data. Analysis of variance computed over the 20-minute session showed no significant differences in mean performance due to the noise conditions and no significant interaction between noise conditions and the level of difficulty of the problems. As expected, difficulty of problems was significant ($p < .01$).

The *F* tests showed that variances computed over the 20-minute session were not significantly different with the D groups in noise as compared with room noise conditions. With the E groups variances differed significantly ($p < .01$).

Mean number of correctly completed problems per 5-minute interval was plotted for each group. There was no evidence of a decrement in performance over time which could be attributed to noise in either the D groups or the E groups.

A test of significance of the differences between variances for successive 5-minute periods under each condition showed no significant differences.

¹ Monitored throughout the session with Sound Level Meter, Type 759, General Radio Company, Cambridge, Massachusetts.

² Nathan Manufacturing Company M3 airchime whistle.

DISCUSSION

The present study shows no effects on mean performance which could be attributed to the noise level. On the other hand, increased noise level increased the variability of performance for the subjects working the easy problems. The fact that this increase in variability occurred with the easy problems but not with the difficult problems does not accord with the literature which reported easier tasks to be less affected by noise than more difficult ones (Broadbent, 1955). In the present study, probably neither of the tasks was particularly difficult for the subjects.

REFERENCES

- BROADBENT, D. E. Noise: Its effect on behaviour. *Roy. Soc. Hlth. J.*, 1955, **75**, 542-545.
- BROADBENT, D. E. Effect of noise on an "intellectual" task. *J. Acoust. Soc. Amer.*, 1958, **30**, 824-827. (a)
- BROADBENT, D. E. *Perception and communication*. New York: Pergamon Press, 1958. (b)
- JERISON, H. J. Effects of noise on human performance. *J. appl. Psychol.*, 1959, **43**, 96-102.
- VITELES, M. S., & SMITH, K. R. An experimental investigation of the effect of change in atmospheric conditions and noise upon performance. *Trans. Amer. Soc. Heat. Vent. Engineers*, 1946, **52**, 167-182.

(Received October 8, 1962)

WORK SATISFACTION AND SCORES ON A PICTURE INTEREST INVENTORY

HAROLD GEIST

Berkeley, California

Men in 6 different occupational groups were given the Geist Picture Interest Inventory (GPII) and the Hoppock Job Satisfaction Survey. Correlations were computed between job satisfaction variables and scores on relevant scales of the GPII. Additional data were gathered in regard to satisfaction and dissatisfaction with work. With the exception of the Clerical and Outdoor scales, the scales of the GPII used in this study appear to be valid, using work satisfaction as a criterion of validity. Median correlations of the other groups range from .209 (social workers) to .866 (artists). "Freedom" and "intellectual stimulation" were the 2 most prominent reasons for satisfaction while "lack of appreciation by colleagues and administrators" and "bad physical working conditions" were most prominent for disliking work.

The whole concept of vocational satisfaction is an oft disputed one. Strong (1955) claims that it is virtually impossible to separate the various facets of satisfaction and that occupational satisfaction is highly inter-related with the physical and social environment surrounding the job. Kuder (1948) and Super (1949) on the other hand believe that occupational satisfaction is the final criterion of the validity of interest measurement. Sarbin and Anderson (1942) related *statements* of vocational satisfaction or dissatisfaction with interest patterns as measured by the Strong Vocational Interest Blank. They found that 82% of the men who expressed dissatisfaction with their current occupations did not have primary interest patterns in the field in which they were employed. They state "adults who complain of occupational dissatisfaction show, in general, measured interest patterns which are not congruent with their present or modal occupation."

Despite the above difference of opinion on work satisfaction and because of the dearth of validity criteria in interest measurement, work satisfaction is still one of the major means of validating interest tests. Both Strong (1955) and more recently Kuder (1948) used this criterion to validate their instruments.

The Geist Picture Interest Inventory (GPII) is a pictorial interest inventory that consists of 129 pictures of occupations and 3 of hobbies. The pictures are presented in triad

form with a forced-choice selection of 1 of the 3 pictures. Scoring categories are the same as the Kuder with the addition of a dramatic scale. The test was standardized on male school children in Grades 8-12, in college, and trade schools in this country, Puerto Rico, and Hawaii. There is also a semiprojective portion of the test where motivating forces behind occupational choice are assessed by asking the testee questions about the pictures of choice. These motivating forces are in the areas of prestige, finance, environment, family, personality, and past experience.

METHOD

Satisfaction and dissatisfaction in the chosen work were ascertained for 271 men in six different occupational groups. These testees were part in a group on which adult norms had previously been obtained for the GPII. The method of obtaining the information was a modified questionnaire of the Hoppock Job Satisfaction Survey (Hoppock, Robinson, & Zlatchin, 1948). The questionnaire used in the current study consisted of five forced-choice sections, two semi-open-end sections relating to things liked best and most about their job, one open-end question relating to feeling about the job in general. There was also one section in which the testee checked how well satisfied he was with his last three jobs. The forced-choice questions were in these areas:

- A. Feelings toward job in terms of liking or disliking.
- B. Time spent feeling satisfied.
- C. Feelings about changing jobs.
- D. Comparison with other people on how they liked their job.
- E. Choice of jobs.

TABLE 1
OCCUPATIONS AND RESPECTIVE SCALE AREAS OF THE SAMPLE

Occupation ^a	Scale area	N	Age range	Median number of years in occupation
1. Clerks	Clerical	37	19-40	6
2. Mathematicians	Computational	71	19-33	7
3. P. E. Teachers	Outdoor	30	22-48	9
4. Artists	Artistic	64	21-26	12
5. Social workers	Social Service	41	28-37	10
6. Scientists (physical)	Scientific	28	26-50	18

Note.—Total N = 271.
^a Most of these people were fairly successful in their chosen occupations. For example, much of the sample was taken from directories of "Who's Who" in the respective occupation.

Each response was scored on the basis of a fixed rating scale (the same number for each category, but different for different categories, e.g., A had 10 possible answers; B, 7; etc.). Each of the testees were also placed in a scale area according to the work they did. For example, the artists were placed in the artistic scale area, the lawyers in persuasive, etc. This is what is termed placing an occupation in a "respective" scale. The author made the arbitrary decision of placing an occupation in a particular scale and it was done on the basis of how close the particular occupation came in the author's judgment to fitting in within the meaning and definition of a particular scale. Occupations and respective scale areas of the sample are shown in Table 1.

The sample was also given the GPII. Correlations were done among each of the forced-choice categories of the questionnaire, and between the forced-choice categories and scores on respective scale. For example, in the case of the clerks, scores on the Clerical scale were correlated with the forced-choice categories; for the artists, scores on the Artistic scale were correlated with the categories; etc.

RESULTS

Table 2 indicates how satisfied the people in a given occupation are with their work as compared with the other parts of the sample.

In general, artists and social workers are satisfied with their work while clerks are not. A mean satisfaction score was computed for each group as a whole (see Table 3).

Table 4 indicates the intercorrelations of the various satisfaction-dissatisfaction categories. The correlations are moderately high indicating that there is a core of satisfaction-dissatisfaction that the questionnaire is measuring, but also that each category measures something distinct.

The correlations of each of the categories with each of the scales for each of the occupations sampled have been computed and are available to the interested reader.¹ Table 5 shows the respective scale intercorrelations. An inspection of Table 5 indicates that all correlations of the five categories of the questionnaire for each occupational group are positively correlated with the respective scale of the GPII with the exception of the clerks and physical education (P. E.) teachers. This means one of two things: either these two groups have high satisfaction and low interest or low satisfaction and high interest. (Thus the clerks expressed high satisfaction with their clerical duties and low inventoried clerical interests or vice versa. The same holds true for P. E. teachers insofar as satisfaction with gym duties and inventoried outdoor interests.) Further inspection of Table 5 indicates that the highest median correlations are among the artists and in descending order scientists, mathematicians, social workers, clerks, and P. E. teachers. All the positive correlations are statistically significant, indicating that with the exception of the Clerical and Outdoor scales, using work satisfaction as a criterion, these scales of the GPII were

¹ The original correlation tables have been deposited with the American Documentation Institute. Order Document No. 7636 from ADI Auxiliary Publications Project, Photoduplication Service, Library of Congress, Washington 25, D. C. Remit in advance \$1.25 for microfilm or \$1.25 for photocopies and make checks payable to: Chief, Photoduplication Service, Library of Congress.

TABLE 2

MEANS AND STANDARD DEVIATIONS OF OCCUPATIONAL GROUPS IN SATISFACTION-DISSATISFACTION CATEGORIES:
ABILITY TO TELL HOW WELL EACH GROUP LIKED THEIR JOB COMPARED WITH OTHER GROUPS

Category	Clerks		Mathema- ticians		P. E. teachers		Artists		Social workers		Scientists		p
	Means	SD	Means	SD	Means	SD	Means	SD	Means	SD	Means	SD	
A. Intensity of feeling like-dislike	7.6	1.7	8.0	1.4	8.5	.58	8.7	1.2	8.9	.78	8.1	.69	.001
B. Amount of time spent feeling satisfied	5.6	1.5	6.0	0.0	6.0	.82	6.3	.58	5.1	2.5	5.9	.38	.001
C. Changing jobs	4.6	1.1	5.1	.69	5.0	.82	6.0	0.0	5.4	1.0	5.7	.76	.001
D. Comparison with other people on liking job	4.6	.89	4.7	.76	3.8	2.5	5.0	0.0	5.1	.33	4.7	.76	.001
E. Choice of jobs	3.0	1.2	2.6	.98	3.3	.96	3.7	.58	3.6	.53	2.4	1.5	.001
N	37		71		30		64		41		28		

valid. However, an inspection of the intercorrelations for all the scales for each of the occupations indicates individual nuances within the respective scales as compared with the other scales for each of the categories. In general, where a respective scale intercorrelation is exceeded by another scale intercorrelation within the same category, the intercorrelations between these scales are usually high (see Geist, 1959, p. 435). For example, for the mathematicians, the respective scale is Scale 8 (Computational). In Category A (intensity of feeling in terms of liking or disliking), the correlation between this category and Scale 8 for the mathematicians is exceeded slightly by Scale 2 (Clerical) and Scale 10 (Social Service) correlations. In Table E,¹ the social workers' respective scale is Scale 10 (Social Service). The correlation between the respective scale (Social Service) and Category A is exceeded slightly by the correlation between Scale 6 (Outdoor) and Category A. Thus, while in some instances the respective scale correlation with a category is exceeded by the correlation(s) of other scales and that same category, in general, these other scales are similar to the respective scale in meaning and content.

The difficulties associated with ascertaining the relationship between job preference and job satisfaction stem from the problem of eliciting job dissatisfaction from people engaged in various kinds of work, particularly in the professional and skilled vocations. Very few are dissatisfied with their jobs and very few want to change jobs. A tally was made of results on a 5-point self-rating scale where the testee was asked to rate himself whether on the last three jobs he was completely dissatisfied, more dissatisfied than satisfied, about half satisfied and half dissatisfied, more satisfied than dissatisfied, and completely satisfied. Only 3 were completely dissatisfied with their jobs, 15 more dissatisfied than satisfied, 15 half and half, and the remainder more satisfied than dissatisfied or completely satisfied. A tally was also made of the three elements that the people liked best about their jobs and the three things they disliked the most. The classifications of these likes and dislikes turned out to be a

TABLE 3
MEAN SATISFACTION SCORE
FOR EACH GROUP

Occupation	Mean satisfaction score
Artists	5.9
Social workers	5.6
Scientists	5.36
P. E. teachers	5.42
Mathematicians	5.38
Clerks	5.1

Herculean job since there were so many varied answers that it was difficult to categorize them. Finally, 21 categories were devised for "liking" and "disliking." The two reasons which the majority in this sample gave for liking their work were "freedom" and "intellectual stimulation." Then in descending order were such items as "associa-

tion with young people," "serving others," "salary," "variety in work," "working time," "challenge," and "association with colleagues." Far down on the list was "prestige," "responsibility," "environment," "location," and "security, which are often thought of as potent reasons for liking one's work. In the "dislikes," the most prominent reasons for disliking work were "lack of appreciation of effort both by administration and colleagues," and "bad working conditions" (poor physical environment, equipment, etc.); another reason which a large number complained about was "some personal aspect of the work," for example, one coach complained he did not like losing in competition. Other reasons in descending popular order were "no room to advance," "salary," "pressure of time," "working hours." Surprisingly few expressed "tension," "location," and "low status," and "politics." One lawyer stated one of the

TABLE 4
INTERCORRELATIONS OF VARIOUS SATISFACTION-DISSATISFACTION CATEGORIES

Category	A Intensity of feeling like-dislike	B Amount of time spent feeling satisfied	C Changing jobs	D Comparison with other people on liking job	E Choice of jobs
A		275	701	415	538
B			291	310	326
C				449	569
D					411
E					

TABLE 5
CORRELATION OF SCORES ON "RESPECTIVE" SCALE WITH
SATISFACTION-DISSATISFACTION CATEGORIES

Occupation	A Intensity of feeling like-dislike	B Amount of time spent feeling satisfied	C Changing jobs	D Comparison with other people on liking job	E Choice of jobs	Median
Clerks	-562	-700	-085	-647	-295	-562
Mathematicians	448	370	243	393	457	393
P. E. teachers	-650	-921	-736	-977	746	-736
Artists	999	866	748	832	866	866
Social workers	262	-197	171	661	209	209
Scientists	324	869	626	283	686	626
Median	293	086	207	338	572	

reasons he disliked his work was because he had to charge for his services.

The results of this study agree in some respects with Strong's (1955) follow up of Stanford students on the reasons for satisfaction or dissatisfaction. Like Strong, the author found that "health" was not an important factor (no one mentioned it in this study). "Criticisms of superior officials" was not mentioned more as a cause for dissatisfaction than "incompetence." "Routine" was not mentioned frequently as it was in Strong's study, probably because of the high caliber of jobs in this project, although Strong's group also was highly loaded in the professional group. Although many of the men in this group lived through the great depression, "security" was not an important factor; "responsibility," "freedom from interference," and "intellectual stimulation" were more im-

portant as reasons for satisfaction than "security."

REFERENCES

- GEIST, H. Manual for Geist Picture Interest Inventory, General Form, Male. *Psychol. Rep. monogr. Suppl.*, 1959, No. 3, 413-438.
- HOPPOCK, R., ROBINSON, H. A., & ZLATCHIN, P. J. Job satisfaction researches of 1946-47. *Occupations*, 1948, 27, 167-175.
- KUDER, G. F. *Examiner manual for the Kuder Preference Record, Personal Form A*, 88-12. Chicago, Ill.: Science Research Associates, 1948.
- SARBIN, T. R., & ANDERSON, H. C. Preliminary study of the relation of measured patterns and occupational dissatisfactions. *Educ. psychol. Measmt.*, 1942, 2, 23-36.
- STRONG, E. K., JR. *Vocational interests 18 years after college*. Minneapolis: Univer. Minnesota Press, 1955.
- SUPER, D. *Appraising vocational fitness by means of psychological tests*. New York: Harper, 1949.

(Received October 30, 1962)

INFLUENCE OF MULTIPLE-CHOICE ANSWER FORM DESIGN ON ANSWER-MARKING PERFORMANCE

IRWIN MILLER AND FRANK J. MINOR¹

International Business Machines Corporation, Endicott, New York

An evaluation of 8 potential new multiple-choice answer forms was conducted. The purpose of the evaluation was to determine whether the new forms differed significantly from a standard IBM form in facilitating the marking of answers. The new forms and the standard IBM form were administered to 4th graders, 8th graders, and college students by means of a timed answer-marking task. 6 of the forms produced significantly lower Total Right and Total Erasures scores than the IBM form. Differences in Total Double Marks were not significant. Females were found to achieve consistently higher mean Total Right scores than males on all answer forms.

This study compared eight experimental answer forms to a standard IBM answer form widely employed with multiple-choice tests. The purpose of the study was to determine whether the experimental forms differed significantly from the standard form in terms of how quickly and accurately respondents could mark designated answers.

All of the experimental forms were of the five-alternative type and all were designed within limitations imposed by projected engineering changes in test scoring equipment. More specifically, the positioning of the response marking areas (hereafter referred to as *slots*) was invariant, as was the color in which the forms were printed (red). However, systematic variations were introduced in the manner in which the slots were organized into groups of five, the manner in which they were labeled, and the type of boundary used to delimit the slots.

Evidence regarding the influence of these changes was deemed essential since a new answer form that caused marking errors or required more time to mark answers could conceivably handicap a person's test performance and invalidate established test norms. To investigate the likelihood of such undesirable effects, the experimental forms and the standard form were administered to three samples of subjects representing different age and educational levels by means of an answer-marking task.

EXPERIMENTAL VARIABLES

The standard answer form which served as a control condition was IBM Form I. T. S. 1000A 155, hereafter referred to as the IBM form. Every answer item in this form consists of five slots, numbered 1, 2, 3, 4, 5 at their upper ends and arranged as a horizontal group (Figure 1A).

In the experimental forms, unlike the IBM form, both the individual slots and the entire answer items (numbered groups of five slots) were oriented parallel with the short edge of the answer form. Within this limitation, four basic configurations were established:

1. Horizontal-Individual (HI): In this configuration (Figure 1B), the five slots that constituted an answer item were presented as a horizontal group and ordered from left-to-right by an individual numeral at the left of each slot.

2. Horizontal-Common (HC): In this configuration (Figure 1C) the slots were also presented as a horizontal group, but were ordered from left-to-right by a set of common numerals positioned midway between every adjacent pair of answer items.

3. Vertical-Individual (VI): In this configuration (Figure 1D) the slots were presented as a vertical group and were ordered from top-to-bottom by an individual numeral at the top of each slot.

4. Vertical-Common (VC): In this configuration (Figure 1E) the slots were also presented as a vertical group, but were ordered from top-to-bottom by a set of com-

¹ Now with IBM Advanced Systems Development Division, Yorktown Heights, New York.

METHOD

Answer-Marking Task

The answer-marking task devised was intended to emphasize the marking process as the primary object of investigation, isolated as far as practicable from additional skill or knowledge requirements. A set of five equivalent booklets was assembled, each intended for administration with a single answer form. Each booklet consisted of 10 pages and a cover. On every page a list of 15 "question numbers" was presented, with an "answer" numeral beside each of the question numbers. The answer numeral, ranging from 1 to 5, was selected by means of random number tables. The task of the subject was to mark (on his answer form) the slots specified by the answer numerals.

The task was complicated slightly by the fact that the questions were not listed consecutively in the test booklet, but in a manner that required the subject to turn to the next page after marking each answer. Thus the first page contained Question Numbers 1, 11, 21 . . . 141; the second page, 2, 12, 22 . . . 142; the third page, 3, 13, 23 . . . 143, etc. After marking the answer for an item listed on page 10, the subject was to return to page 1 for the next item and proceed through the booklet again.

This arrangement of the question numbers within the booklets was adopted deliberately because it provided no opportunity for subjects to memorize several answers at a glance and mark them at one time. Hence, each answer marked on a form entailed a distinct perceptual-motor search.

Subjects

Elementary school children² were represented by four Grade 4 classes, comprising a total of 102 pupils, 52 males and 50 females. The children ranged in age from 8 to 11, with a mean age of 9.3 years.

An older public school sample was represented by four Grade 8 classes, a total of 132 pupils. The Grade 8 pupils consisted of 58 males and 74 females, and ranged in age from 13 to 17 with a mean age of 13.6 years.

Freshman and sophomore students enrolled at a liberal arts college (Harpur College) participated in the study as paid volunteers. A total of 49 college students, 24 males and 25 females, took part in three classroom sessions. The college students ranged in age from 17 to 20 with a mean age of 18.3 years.

Administration Procedure

To demonstrate how the various forms were to be marked, the experimenter made use of a set of placards, each exhibiting a greatly enlarged drawing of answer item Number 1. Before each form was administered, the appropriate placard for that form

▪ The cooperation of the Binghamton public school system in making available both the Grade 4 and Grade 8 pupils is gratefully acknowledged.

(A)

	1	2	3	4	5
1	11	11	11	11	11
	11	11	11	11	11

(B)

	1	2	3	4	5
1	11	11	11	11	11
	11	11	11	11	11

(C)

	1	2	3	4	5
1	11	11	11	11	11
	11	11	11	11	11

(D)

	1	2	3	4	5
1	11	11	11	11	11
	11	11	11	11	11

(E)

	1	2	3	4	5
1	11	11	11	11	11
	11	11	11	11	11

(F)

	1	2	3	4	5
1	11	11	11	11	11
	11	11	11	11	11

FIG. 1. Answer items.

mon numerals positioned midway between every pair of adjacent answer items.

In one respect the four experimental forms described above resembled the IBM form: namely, each slot consisted of a pair of dashed lines. An additional set of four experimental forms was designed in which each slot was a closed rectangle. In all other aspects these additional forms corresponded completely to the four answer forms with the dashed line slots. As an illustrative example, the solid line version of the HI configuration is shown in Figure 1F.

As noted above all eight experimental forms were printed in red ink instead of the blue ink typical of IBM forms. In the present study the IBM form was also printed in red ink to limit the investigation to configurational effects only, unconfounded with possible color effects.

was shown and then marked with a black wax pencil to demonstrate the proper way of indicating the answer to question Number 1, as given in the test booklet.

After completing the first three questions as practice items, the subjects were provided with a complete explanation of how the booklet was numbered. They were then instructed that, at the experimenter's signal, they were to continue to "mark as carefully and as quickly as you can."

Each form was administered for a 10-minute period in Grade 4 and Grade 8 classes. A 7-minute period was used for the college students since a pilot study indicated that 10 minutes would allow most of the college students to complete most of the forms. A total of five different forms was administered to every subject. Between the third and fourth form presented, a rest period of about 10 minutes was introduced.

Experimental Design

For each of the three subject samples an experimental design involving repeated measurements of independent classroom groups was carried out. In the course of a single classroom session, each group was administered the IBM form, an HI form, an HC form, a VI form, and a VC form. Aside from the IBM form, only the dashed line or the solid line forms (not both) were administered to the subjects in a given classroom group. Partial counterbalancing of form presentation order was accomplished by reversing presentation orders within each subject sample.

RESULTS

Three dependent variables were used as criteria of answer-marking performance:

- 1. Total Right: The total number of items with the correct alternative marked and no other alternatives marked.
- 2. Total Erasures: The total number of items with evidence of one or more erasures.
- 3. Total Double Marks: The total number of items with two or more alternatives marked and no evidence of attempted erasure.

Total Right

Table 1 presents the mean Total Right scores for the IBM form and the HI, HC, VI, and VC forms. The means of the experimental forms are based on the pooling together of the groups that were administered forms in a dashed line version and the groups administered the solid line forms. As Table 1 makes clear, striking differences occurred among the forms in all three subject samples.

For each row of subjects presented in

TABLE 1
TOTAL RIGHT: MEAN SCORES BY FORM,
EDUCATIONAL LEVEL, AND SEX

Subjects	Form				
	IBM	HI	HC	VI	VC
Grade 4 males	75.6	73.4	71.0	67.6	62.0
Grade 4 females	82.3	78.7	72.2	70.8	62.8
Grade 8 males	114.4	110.9	103.0	102.7	91.6
Grade 8 females	120.0	119.5	114.7	111.0	101.4
College males ^a	125.0	124.8	126.1	117.1	104.0
College females ^a	137.4	137.7	135.7	125.6	117.0

^a Seven minutes.

Table 1, an analysis of variance was conducted to test differences among forms, classroom groups, and the Forms × Groups interactions. The outcomes of these analyses are summarized in Table 2. In every instance differences among forms were significant and the Forms × Groups interactions were also significant. The differences among groups were not significant in the college sample but did attain significance in Grade 8 males and Grade 4 females.

To identify more specifically the forms that had produced the overall significant differences, additional tests were made in accordance with the procedure developed by Scheffé (1953) for the individual comparisons of means in the analysis of variance. The outcomes of these tests are summarized in Table 3.

The HI forms did not differ significantly from the IBM form in regard to Total Right

TABLE 2
TOTAL RIGHT: SIGNIFICANCE OF DIFFERENCES
IN MEAN SCORES AMONG FORMS AND GROUPS

Subjects	Differ- ences among forms	Differ- ences among groups	Forms × Groups inter- action
Grade 4 males	.01	<i>ns</i>	.01
Grade 4 females	.01	.01	.01
Grade 8 males	.01	.05	.01
Grade 8 females	.01	<i>ns</i>	.01
College males ^a	.01	<i>ns</i>	.01
College females ^a	.01	<i>ns</i>	.01

^a Seven minutes.

in any subject sample. The HC forms did not differ significantly from the IBM form in the college sample, but did yield significantly lower scores in all public school comparisons. For the VI forms and the VC forms the Total Right scores proved significantly lower than the IBM form in every subject sample.

Practice Effect. In Figure 2 the mean Total Right scores are plotted as a function of order of form presentation. The predominant characteristic of these curves is a tendency for the totals to increase as successive forms are administered. Since the presentation orders were partially counterbalanced, however, the curves would not be expected to rise unless there was a practice effect; a conclusion supported by the significant Forms

TABLE 3

TOTAL RIGHT: DIFFERENCES BETWEEN MEAN SCORE OF IBM FORM AND MEANS OF HI, HC, VI, AND VC FORMS THAT PROVED SIGNIFICANT

Subjects	IBM vs. HI	IBM vs. HC	IBM vs. VI	IBM vs. VC
Grade 4 males	<i>ns</i>	4.6	8.0	13.6
Grade 4 females	<i>ns</i>	10.1	11.5	19.5
Grade 8 males	<i>ns</i>	11.4	11.7	22.8
Grade 8 females	<i>ns</i>	5.7	9.0	18.6
College males ^a	<i>ns</i>	<i>ns</i>	7.9	21.0
College females ^a	<i>ns</i>	<i>ns</i>	11.8	20.4

^a Seven minutes.

× Groups interactions reported in Table 2. All subject samples show large improvement between the first and second form and then tend to reach a plateau. Evidently the public school groups continued to improve after the fourth form, whereas the college students did not.

A very interesting finding is the consistent, often substantial superiority of females in all three subject samples (see also Table 1). Even with magnitude of difference ignored, the probability of this sex difference occurring by chance is 2⁻¹⁵.

Dashed Line Forms and Solid Line Forms. To investigate the influence of the type of line used to delimit the marking slots, additional tests were made in each subject

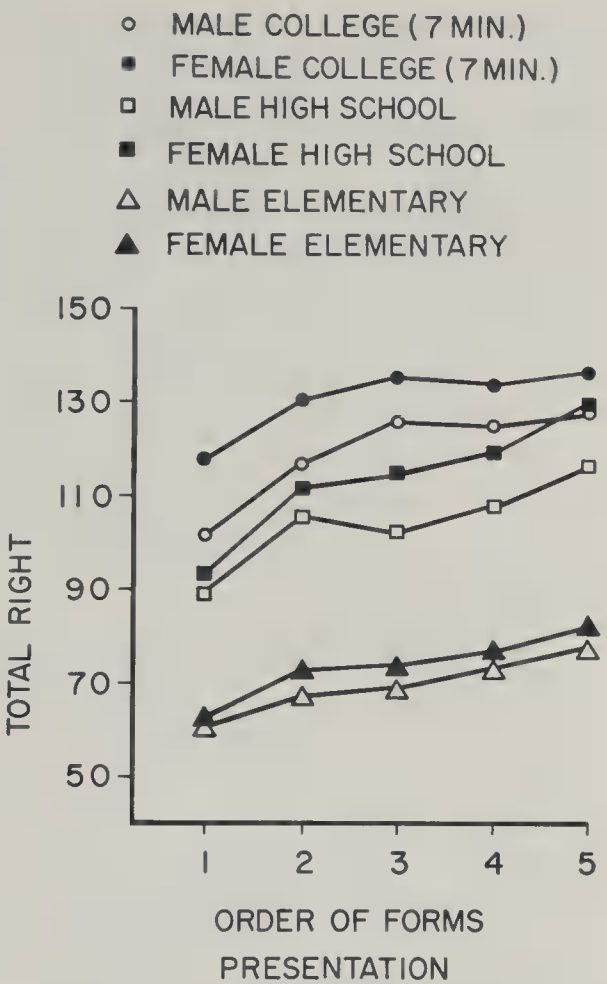


FIG. 2. Total Right items. (Means of Total Right as a function of order in which forms are presented.)

sample of the differences between the groups given the dashed line forms and the groups given the solid line forms. The results of these tests, made in accordance with the Scheffé procedure, indicated that the dashed line and solid line forms did not differ significantly

TABLE 4

TOTAL DOUBLE MARKS AND TOTAL ERASURES: MEAN SCORES FOR ALL EXPERIMENTAL FORMS AND IBM FORM COMBINED

Subjects	Double Marks	Erasures
Grade 4 males	.56	1.40
Grade 4 females	.42	1.68
Grade 8 males	.21	1.32
Grade 8 females	.22	1.54
College males ^a	.65	1.11
College females ^a	.29	1.40

^a Seven minutes.

($p > .05$) regardless of subject sample or sex.

Total Erasures and Total Double Marks

The means of Total Double Marks and Total Erasures for all forms combined (IBM and experimental) are presented in Table 4. For convenience in computing the Total Erasure values, a maximum score of 9 was assigned in the few instances where an individual's Total Erasures exceeded 9 (this occurred three times in the Grade 4 sample, four times in the Grade 8 sample, and twice in the college sample). As Table 4 indicates, the means for Total Double Marks did not exceed .65 of an item; the mean Total Erasures, 1.68 items.

Since the frequency data for erasures and double marks formed extremely skewed distributions, making the normal distribution assumption untenable, the Friedman χ^2 test (1937), a nonparametric test, was applied to test the differences among forms within each subject sample. The outcome of the χ^2 tests failed to indicate any significant differences in Total Double Marks among the forms ($p > .05$). In regard to Total Erasures no significant differences were found in the college sample or Grade 8 sample ($p > .05$); however, significant differences did occur among the Grade 4 males and females ($p < .05$), with the IBM form yielding the highest totals, 1.90 items and 2.34 items, respectively.

To determine specifically which forms had produced the Total Erasure differences in the Grade 4 sample, χ^2 tests were conducted comparing the HI, HC, VI, and VC forms

with the IBM form individually. The results of these tests are summarized in Table 5 where the mean differences in Total Erasures are reported for those individual comparisons that proved significant. Only the HI form did not differ significantly from the IBM form for both males and females. Although the interpretation of these findings is hardly clear-cut, it may be noted that the non-significant difference in the comparison of the IBM form and the HI form is consistent with the findings noted in Table 3, and that all differences are very small.

DISCUSSION

Of the three answer-marking criteria employed—Total Right, Total Erasures, and Total Double Marks—only the Total Right scores reflected differences between the IBM form and the experimental forms that were both statistically significant and of practical import. Since the type of line used to delimit slots had no significant effect on Total Right scores, the differences among the forms can be attributed to the configurational factors involved in the grouping and the labeling of the slots. Some inferences can be made as to the relative importance of these factors from the pattern of significant differences in Table 3.

Evidently vertical grouping of slots was the factor most detrimental to answer-marking performance since the two vertical configurations, VI and VC, consistently produced the lowest Total Right scores, and handicapped the college students as well as the public school pupils. Aside from the adverse effect of vertical grouping, Table 3 strongly implies that failure to label every slot with an individual numeral also had an adverse effect on answer-marking performance. This implication is borne out by the fact that scores on the HC forms were significantly lower than the IBM form in the public school samples while the HI scores did not differ significantly for those subjects. In addition it should be noted that the VC scores were consistently lower than the VI scores in all three samples (Table 1).

As a matter of practical interpretation, the significant differences in Total Right scores

TABLE 5

TOTAL ERASURES: DIFFERENCE BETWEEN MEAN OF IBM FORM AND MEANS OF HI, HC, VI, AND VC FORMS THAT PROVED SIGNIFICANT FOR ELEMENTARY SCHOOL SUBJECTS

Subjects	IBM	IBM	IBM	IBM
	vs. HI	vs. HC	vs. VI	vs. VC
Males	ns	.77	.67	.57
Females	ns	.86	ns	1.18

provide evidence that the HC, VI, and VC forms are unsuitable for use with tests standardized on IBM forms. On the other hand, it would be premature to conclude that the HI forms may be accepted for general use on a basis of the present results alone. Additional research on the possible influence of color is necessary, for example, since the IBM form used as a control condition was specially printed in red ink rather than the conventional blue ink. More fundamentally, however, it must be recognized that although the answer-marking task appears to be a useful screening device, it is dissimilar from an actual test administration. In particular, since the absence of substantive test content and the manipulation of the test booklet are contrary to ordinary testing procedure, a validating field study incorporating recognized standard tests is in order.

The finding that females achieved higher Total Right scores than males on every form studied, including the IBM form, raises a provocative question. It is conceivable that this sex difference is intrinsic to the experimental marking task; for example, the females may have turned the pages with less fumbling after marking each item and thereby secured an advantage not available in conventional test taking. On the other hand, if

the sex difference observed can be shown to be independent of the special features of the answer-marking task, it may be inferred that females habitually receive a "free" score increment on tests administered with standard answer forms regardless of test content.

Pyle (1925, Ch. 4) reported a consistent difference favoring females in a digit-letter substitution task, based on subjects ranging in age from 8 to 18. The substitution task differed from the present answer-marking task in several particulars, perhaps most importantly in that it utilized a key, a fixed association scheme of given digits and letters, which the subject could learn to his advantage. The present findings suggest that sex differences in such substitution tasks may be less a function of associative learning than of fundamental perceptual motor skills.

REFERENCES

- FRIEDMAN, M. The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *J. Amer. Statist. Ass.*, 1937, **32**, 675-701.
PYLE, W. H. *Nature and development of learning capacity*. Baltimore, Md.: Warwick & York, 1925.
SCHEFFÉ, H. A method for judging all contrasts in the analysis of variance. *Biometrika*, 1953, **40**, 87-104.

(Received November 8, 1962)

INFLUENCE OF SIMULTANEOUS VARIATION IN SIZE OF TYPE, WIDTH OF LINE, AND LEADING FOR NEWSPAPER TYPE¹

MILES A. TINKER

Santa Barbara, California

A speed of reading technique was employed with 820 readers to determine the effect on legibility of simultaneous variation of type size, line width, and leading for Excelsior newspaper type. 9 typographical variations were compared with text set in 8-point type in a 12-pica line width with 2-point leading. Results revealed that 7-, 8-, and 9-point type in a 12-pica line width with 2-point leading were read most rapidly and equally fast. But text in relatively long lines, very short lines, and small type size, or combinations of these with little or no leading were read significantly slower than the standard. Judgments for relative legibility and pleasingness revealed a reader preference for 8- or 9-point type with 2-point leading in a line width of 12 picas (12 or 18 picas for 9-point type). Text in relatively long or short lines, small type size, and no leading received low ratings.

Investigations of the legibility of typographical arrangements by Paterson and Tinker (1940) have shown that determination of an optimal printing arrangement is possible only when type size, line width, and leading are varied together. This holds both for book and for newspaper printing. After several preliminary investigations of newspaper printing (see Tinker, 1963), the present experiment was carried out. The problem is to discover the effect on legibility of simultaneous variation of type size, line width, and leading for newspaper type.

METHOD

A speed of reading technique was employed. The reading material consisted of Forms A and B of the Chapman-Cook Speed of Reading Test. This test measures speed virtually as a single variable since comprehension is constant with 99.7% accuracy. (For methodology in use of test see Tinker & Paterson, 1936.) In all comparisons, Form A was printed in 8-point type in a 12-pica line width with 2-point leading as a standard. Form B, which was read after Form A, was set in the typographical arrangements shown in Table 1; that is, type size, line width, and leading varied from group to group. All printing was in Excelsior type face on newspaper stock.

Ten groups of 82 high school seniors each were tested. In Group I, the control group, the typography was identical in Forms A and B. Thus a correction

can be made in Groups II through X (the experimental groups) for whatever deviation occurs between Forms A and B of the control group. In Group II, and each of the successive groups, the text in 8-point type in a 12-pica line width with 2-point leading was compared with text with the variations in type size, line width, and leading shown in Table 1. These comparisons will reveal which arrangements produce the faster reading; that is, the more legible (or readable) print.

RESULTS

The detailed statistical results are shown in Table 1. It will be noted that 7-, 8-, and 9-point type in a 12-pica line width with 2-point leading are read equally fast. The 9-point, 18-pica with 2-point leading arrangement, and the 7-point, 12-pica with 1-point leading setup are read nearly as fast.

In contrast with the above are the typographical arrangements that retard speed of reading significantly: relatively long lines (Groups IV, V), very short lines (Groups VI, IX), small type size (Groups IX, X), or combinations of these with little or no leading.

Reading judgments of legibility for samples of the typographical arrangements used above are illuminating. Results are given in Table 2. The texts read fastest tend to be ranked high and those read slowest, ranked low. In general, readers prefer 8- and 9-point type in 12- or 18-pica lines with 2-point leading (highest four ranks). They dislike small

¹ The writer is grateful to the University of Minnesota Graduate School for a research grant to finance this study. Data for this study were collected jointly by the writer and Donald G. Paterson.

TABLE 1
INFLUENCE OF SIMULTANEOUS VARIATION IN SIZE OF TYPE, WIDTH OF LINE,
AND LEADING FOR NEWSPAPER TYPE

Test group	Comparison	Mean	P.E. dist.	P.E. mean	Difference between means in:		P.E. diff.	<i>r</i>	D
					Para-graphs ^a	Per-cent			P.E. diff.
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
I	A 8-pt., 12-pica, 2-pt. leading	13.57	2.03	.22	.00	.00	.00	.73	.00
	B 8-pt., 12-pica, 2-pt. leading	14.20	1.73	.19					
II	A 8-pt., 12-pica, 2-pt. leading	14.32	2.62	.29	— .27	1.90	.15	.86	1.78
	B 9-pt., 12-pica, 2-pt. leading	14.68	2.53	.28					
III	A 8-pt., 12-pica, 2-pt. leading	13.94	2.60	.29	— .37	2.65	.17	.81	2.17
	B 9-pt., 18-pica, 2-pt. leading	14.20	2.33	.26					
IV	A 8-pt., 12-pica, 2-pt. leading	14.05	2.44	.27	—2.02	14.40	.16	.80	12.50
	B 9-pt., 43-pica, solid	12.66	2.07	.23					
V	A 8-pt., 12-pica, 2-pt. leading	14.52	2.65	.29	— .89	6.13	.16	.84	5.63
	B 8-pt., 18-pica, 2-pt. leading	14.26	2.46	.27					
VI	A 8-pt., 12-pica, 2-pt. leading	14.39	2.71	.30	—1.41	9.80	.16	.85	8.90
	B 7-pt., 6-pica, solid	13.61	2.16	.24					
VII	A 8-pt., 12-pica, 2-pt. leading	13.98	2.45	.27	— .27	1.93	.14	.86	1.97
	B 7-pt., 12-pica, 1-pt. leading	14.34	2.05	.23					
VIII	A 8-pt., 12-pica, 2-pt. leading	14.23	2.51	.28	— .01	.07	.17	.81	.06
	B 7-pt., 12-pica, 2-pt. leading	14.85	2.37	.26					
IX	A 8-pt., 12-pica, 2-pt. leading	14.20	2.44	.27	—1.78	12.54	.15	.82	11.59
	B 6-pt., 6-pica, solid	13.05	2.04	.22					
X	A 8-pt., 12-pica, 2-pt. leading	15.01	3.14	.34	— .98	6.53	.18	.85	5.39
	B 6-pt., 12-pica, 2-pt. leading	14.66	2.46	.27					

Note.—Differences given are for the mean score on Form A, 8-point, 12-pica line width with 2-point leading minus the mean score on Form B printed as indicated in comparison. All forms were printed in Excelsior type face on newsprint paper stock. In each test group $N = 82$ high school seniors.

^a The differences in Column 6 are "corrected" by the amount of the difference between the mean scores of Form A and Form B of Test Group I which serves as a control group. The "correction" amounts to .63 paragraphs for each test group comparison.

type, long and short lines, and material set solid (four lowest ranks).

The ratings for pleasingness were obtained from a different group of college students. The results are given in Table 3. Ranks for pleasingness are almost identical with those

for apparent legibility and both are in close agreement with the speed of reading scores. As pointed out by Tinker and Paterson (1942), judged legibility may be accepted as equivalent to pleasingness. And since both tend to approximate speed of reading scores,

TABLE 2
VARIATIONS IN SIZE OF TYPE, WIDTH OF LINE, AND LEADING FOR NEWSPAPER TYPE,
RANKED ACCORDING TO READER OPINIONS OF RELATIVE LEGIBILITY

Type variation	Average rank	<i>SD</i>	Rank order
8-pt., 12-pica, 2-pt. leading	3.32	1.41	3
9-pt., 12-pica, 2-pt. leading	2.62	1.59	2
9-pt., 18-pica, 2-pt. leading	2.25	1.52	1
9-pt., 43-pica, solid	6.31	2.15	7
8-pt., 18-pica, 2-pt., leading	3.36	1.49	4
7-pt., 6-pica, solid	8.60	1.26	9
7-pt., 12-pica, 1-pt. leading	6.21	1.44	6
7-pt., 12-pica, 2-pt. leading	5.17	1.34	5
6-pt., 6-pica, solid	9.60	1.24	10
6-pt., 12-pica, 2-pt. leading	7.55	1.32	8

Note.— $N = 180$ college students.

TABLE 3
VARIATIONS IN SIZE OF TYPE, WIDTH OF LINE, AND LEADING FOR NEWSPAPER TYPE
RANKED ACCORDING TO READER OPINIONS OF PLEASINGNESS

Type variation	Average rank	SD	Rank order
8-pt., 12-pica, 2-pt. leading	2.98	1.35	3
9-pt., 12-pica, 2-pt. leading	2.66	1.65	1.5
9-pt., 18-pica, 2-pt. leading	2.66	1.73	1.5
9-pt., 43-pica, solid	6.87	2.04	7
8-pt., 18-pica, 2-pt. leading	3.44	1.52	4
7-pt., 6-pica, solid	8.79	.76	9
7-pt., 12-pica, 1-pt. leading	5.92	1.48	6
7-pt., 12-pica, 2-pt. leading	4.73	1.50	5
6-pt., 6-pica, solid	9.76	.50	10
6-pt., 12-pica, 2-pt. leading	7.20	1.43	8

Note.—N = 180 college students.

newspaper publishers can achieve both a highly legible and pleasing typography as well as one that readers consider more legible by employing 8- or 9-point type, 2-point leading, and either 12- or 18-pica (presumably 12–18 pica) line width.

REFERENCES

PATERSON, D. G., & TINKER, M. A. *How to make type readable*. New York: Harper, 1940.

TINKER, M. A., & PATERSON, D. G. Studies of typographical factors influencing speed of reading: XIII. Methodological considerations. *J. appl. Psychol.*, 1936, 20, 132–145.
TINKER, M. A., & PATERSON, D. G. Reader preferences and typography. *J. appl. Psychol.*, 1942, 26, 38–40.
TINKER, M. A. *Legibility of print*. Ames: Iowa State Univer. Press, 1963.

(Received November 19, 1962)

PAYMENT HISTORY AS A PREDICTOR OF CREDIT RISK

HOWARD W. HASSLER

Security First National Bank, Los Angeles, California

JAMES H. MYERS AND MAURICE SELDIN

University of Southern California

A numerical scoring system to predict credit payment was developed from the payment records of customers of a large Los Angeles department store. The following types of items were used: number of payments of less than amount due, number of payments missed completely, number of collection notices previously sent, etc. Results showed good ($r_{b1} = .71$) ability to discriminate between potentially good accounts and those likely to require special collection efforts.

Previous studies (Durand, 1941; McGrath, 1960; Myers & Cordner, 1957; Wolbers, 1949) have shown that numerical scoring systems, based upon the weighted application blank approach, are valid predictors of credit payment at the retail level. These studies have utilized personal history information from the credit application form to determine whether or not an applicant should be granted credit.

A closely related problem in retail stores is that of determining, once the applicant has been granted credit, how much *additional* credit should be allowed or how high the account balance should be allowed to go. If the account has been open and active for a number of months or years, it seems better to base these decisions on the account *payment history*, rather than on the biographic information from the original application form, since the latter gets out of date and cannot be kept current easily.

The present study involves the development of a numerical scoring system, based upon payment history, to predict later account delinquency among credit customers of a large Los Angeles department store chain.

METHOD

Management was interested in discriminating between two types of accounts: (a) "open"—those which have a history of regular payment and require no special collection effort, and (b) "special handling"—those which require intensive collection efforts, due to lack of payments, inconsistent payments or failure to keep the account within the agreed upper credit limit.

If accounts which would later require special collection efforts could be identified at an early date, steps could be taken to limit the account balance or to maintain a more rigorous program of collection notices for these accounts.

To provide equal "exposure" or opportunity to default, only accounts open and reasonably active for 15 months prior to the time of the study were used. It was felt this was the earliest that sufficient payment history could be obtained on both types of accounts and that, if successful at this time period, the study could easily be extended to longer time periods at a later date. For accounts open more than 15 months, only the last 15 months of payment history were studied.

Six hundred accounts were randomly drawn from open and special handling files—300 of each type. Analyses of predictor items were done using 200 accounts from each group; the remaining 100 accounts in each group served as cross-validation samples.

The following types of payment history items were obtained from the account ledger card for each of the 600 cases: number of months no payment made (when a payment was due), number of months payment made in amount less than due, number of months credit limit had previously been exceeded, number of collection notices sent, etc. A chi square analysis was made of each predictor item, and items discriminating beyond the .005 level were used in a scoring key which was applied to the 200 cross-validation cases. Table 1 shows the complete list of all payment history items studied and the chi square probability levels for each.

RESULTS AND DISCUSSION

A scoring key was developed from the nine items significant at the .005 level or beyond. Equal weights were used for each item, since previous studies (McGrath, 1960; Myers, 1963) had shown no advantage in using the more cumbersome chi square weights. Favor-

TABLE 1
CHI SQUARE PROBABILITY LEVELS FOR PAYMENT HISTORY ITEMS STUDIED

Item		χ^2	df	p
1.	Number of months payment made in amount due	81.54	7	<.005
2.	Number of months no payment made	96.99	5	<.005
3.	Number of months payment made in amount less than due	89.94	5	<.005
4.	Number of months payment made in amount more than due	21.36	5	<.005
5.	Number of months initial account credit limit exceeded by \$100.00 or more	27.13	3	<.005
6.	Number of months initial account credit limit exceeded by \$50.00 or more	37.86	4	<.005
7.	Number of months first collection notice sent	86.42	3	<.005
8.	Number of months second collection notice sent	9.29	1	<.005
9.	Number of months other collection notices sent	104.61	4	<.005
10.	Ratio of highest balance to beginning balance	5.42	7	.60
11.	Ratio of highest balance to average balance	3.09	4	.45
12.	Trend (up or down) of account balance	8.68	3	.04
13.	Pattern of account balance—each balance related to previous balance	19.74	9	.02
14.	Pattern of account balance—each balance related to initial balance	11.79	9	.25

TABLE 2
SCORES OF OPEN AND SPECIAL HANDLING ACCOUNTS: CROSS-VALIDATION GROUPS

Score	Open accounts		Special handling	
	Frequency	Cumulative percent	Frequency	Cumulative percent
18	0	100	0	100
17	24	100	3	100
16	17	76	3	97
15	8	59	2	94
14	8	51	2	92
13	7	43	2	90
12	7	36	6	88
11	6	29	7	82
10	3	23	6	75
9	4	20	9	69
8	8	16	12	60
7	5	8	17	48
6	1	3	15	31
5	1	2	7	16
4	0	1	3	9
3	1	1	5	6
2	0	0	1	1
1	0	0	0	0
0	0	0	0	0
Total	100		100	

able categories for each item were weighted 2, neutral categories 1, and unfavorable categories 0.

Table 2 shows the result of applying the scoring key to the 200 holdout cases. The biserial correlation of .71 is significant well beyond the .001 level and indicates the clear ability of a payment history scoring system to discriminate between potentially good accounts and those which will need special collection efforts.

Armed with this simple scoring device, an inexperienced clerk can rapidly "score" the ledger card of a customer applying for additional credit to determine the degree of risk which might be involved. For example, a customer scoring 6 or below would be likely to require special collection efforts, since approximately 33% of special handling cases score at or below this point, whereas only 3% of good accounts do. Such a customer should be held to lower credit amounts and could also be informed that failure to improve

upon his payment record might result in further restrictions in his upper credit limit.

Of course, this particular scale applies *only* to accounts open and active for exactly 15 months, but it is a simple matter to adapt it to accounts open for more than this period of time, by scoring only the payment history during the last 15 months.

REFERENCES

- DURAND, D. Risk elements in consumer instalment financing, Report No. 8, 1941, National Bureau of Economic Research, New York.
- MCGRATH, J. J. Improving credit evaluation with a weighted application blank. *J. appl. Psychol.*, 1960, **44**, 325-328.
- MYERS, J. H., & CORDNER, W. C. Increase credit operation profits. *Credit World*, 1957 (Feb.), 12-13.
- MYERS, J. H. Predicting credit risk with a numerical scoring system. *J. appl. Psychol.*, 1963, **47**, 348-352.
- WOLBERS, H. L. The use of the biographical data blank in predicting good and potentially poor credit risks. Unpublished master's thesis, University of Southern California, 1949.

(Received November 26, 1962)

JOB ATTITUDES IN MANAGEMENT:

IV. PERCEIVED DEFICIENCIES IN NEED FULFILLMENT AS A FUNCTION OF SIZE OF COMPANY¹

LYMAN W. PORTER²

University of California, Berkeley

This study was concerned with the relationship between size of organization and perceptions of need satisfaction and need importance in management jobs. A questionnaire provided data from a nationwide sample of 1916 managers. Results showed that at lower levels of management small company managers were more satisfied than large company managers, but at higher levels of management large company managers were more satisfied than small company managers. Size of company had little relationship to the other attitude variable, perceptions of the importance of various needs.

Previous articles in this series on job attitudes in management have examined the effects of both vertical level of position and line-staff type of job on perceived satisfaction and importance of various psychological needs (Porter, 1962, 1963a, 1963b). The present study is an investigation of the effect of size of organization on perceptions of need satisfaction and need importance.

Size, as a variable affecting job attitudes, has been studied mostly in connection with different sized working groups within the same organization. This research has been concentrated almost exclusively on determining the optimum size of work units within relatively large organizations that will bring about maximum morale and job satisfaction. The results of such investigations tend to come to the same conclusion: smaller work units are associated with higher morale and greater job satisfaction. For example, Worthy (1950), in his articles on organization structure and morale, stated: "Our researches demonstrate that mere size is unquestionably

one of the most important factors in determining the quality of employee relationships: the smaller the unit the higher the morale and vice versa [p. 172]." Viteles (1953), in his summary of studies concerned with the influence of the size variable, concluded: "The size of work group affects output and attitudes, which both tend to be better in smaller-sized groups [p. 146]." A similar summary of the literature was made by Strauss and Sayles (1960): "Many studies have shown that employee morale is higher in small groups than in large ones [p. 376]." Thus, the evidence is quite consistent in support of the notion that the smaller the *work group* of *nonmanagement employees*, the better the job attitudes.

Despite the evidence cited above concerning the relationship of size and job attitudes, two possible limitations on generalizations from these data should be noted. First, the units of employees studied in most of these investigations were subparts of the same organization, rather than different organizations. It has not yet been adequately demonstrated that smaller organizations produce more favorable employee attitudes than do larger organizations. When different sized groups are studied within the same organization, the possible advantages that might accrue to the larger units are probably greatly attenuated. That is, since all units are working under the same corporate policies, the larger groups do not tend to be able to provide special benefits for their members

¹ This study was carried out as part of the research program of the Institute of Industrial Relations, University of California, Berkeley. It was started while the author was a Ford Foundation Faculty Research Fellow. The Institute of Social Sciences at the University of California and the American Management Association contributed to the support of the research assistance, and the Computer Center of the University provided facilities for data computations.

² The author is indebted to Mildred Henry, Larry Stewart, and Robert Andrews for assistance in tabulation of the data.

that might be expected because of the potentially greater influence of larger versus smaller groups. When large organizations are compared with small organizations, the picture may be considerably different from that for intraorganization comparisons. If separate organizations of different size are compared with each other, then it may be possible to show that greater size provides some advantages for individuals working in large corporations that are less prevalent for individuals working in small companies.

The second limitation on generalizations from the previous findings on the relation of size of group to job attitudes concerns the fact that most of the past research was carried out on the nonmanagement or worker level of organizations. There are good reasons for presuming that organizational level might have an interaction effect on size in relation to job attitudes. For example, a worker at the bottom of a large organization has a much larger superstructure of organization levels and of sheer numbers of people above him than does a similar worker in a small company. In effect, the worker in the large company has more bosses above him and has less absolute influence on his work environment than does the worker in the small company. However, at the other end of the hierarchy—top management—the picture should be reversed. A top manager in a large company controls or “bosses” more people than a top manager in a smaller organization, and hence has (or should have) more absolute influence in the work situation. To the extent that this analysis of the interaction of size of organization and level of position within the organizational hierarchy is correct, it would lead to the following hypothesis: the higher the organizational level, the relatively more favorable will be the job attitudes of individuals in large organizations compared with those of individuals in small organizations.

The purposes of the present study are two: (a) to compare the perceived need fulfillment deficiencies of managers in large organizations versus those of managers in small organizations; and (b) to make such comparisons at different positions in the managerial hierarchy, so as to determine whether there is an inter-

action between size of organization and level of position in affecting job attitudes. Brief attention will also be paid to the dependent variable of perceived importance of various needs.

METHODS

Questionnaire Instrument

The questionnaire used to collect the data for this study was the same as that described in previous articles (Porter, 1961, 1962). The results were based on 13 items in the questionnaire that dealt with needs classifiable on the basis of a Maslow-type hierarchy of prepotency. A sample item as it appeared in the questionnaire is as follows:

The *feeling of security* in my management position:

- (a) How much is there now?
(min) 1 2 3 4 5 6 7 (max)
- (b) How much should there be?
(min) 1 2 3 4 5 6 7 (max)
- (c) How important is this to me?
(min) 1 2 3 4 5 6 7 (max)

Data on perceived deficiencies in need fulfillment were derived from the answers to Parts a and b of each item, as is explained in the Results section. Data on the importance of each type of need to the respondent were based on answers to Part c of each item.

Categories of Needs and Specific Items

Listed below are the specific items and the five categories of needs studied in this investigation. The items were listed randomly in the questionnaire, but are listed here in a systematic fashion according to their theoretical assignment to respective need categories. This categorization system was designed to adhere closely to Maslow's classification scheme based on the prepotency of needs (Maslow, 1954), although the present system is not identical with it. The system used here has been described in detail in a previous paper (Porter, 1961). The need categories and their respective items follow:

I. Security needs

- 1. The *feeling of security* in my management position

II. Social needs

- 1. The *opportunity*, in my management position, to give help to other people
- 2. The *opportunity to develop close friendships* in my management position

III. Esteem needs

- 1. The *feeling of self-esteem* a person gets from being in my management position
- 2. The *prestige* of my management position inside the company (that is, the regard received from others in the company)

- 3. The *prestige* of my management position *outside* the company (that is, the regard received from others *not* in the company)

IV. Autonomy needs

- 1. The *authority* connected with my management position
- 2. The *opportunity for independent thought and action* in my management position
- 3. The *opportunity*, in my management position, for *participation in the setting of goals*
- 4. The *opportunity*, in my management position, for *participation in the determination of methods and procedures*

V. Self-Actualization needs

- 1. The *opportunity for personal growth and development* in my management position
- 2. The *feeling of self-fulfillment* a person gets from being in my management position (that is, the feeling of being able to use one's own unique capabilities, realizing one's potentialities)
- 3. The *feeling of worthwhile accomplishment* in my management position

Procedure and Sample

The sample of 1,916 managers on which the data of this study were based, was obtained by mailing the questionnaire to several thousand individuals in management positions in firms located throughout the country. These individuals represented a random 10% sample of the American Management Association and a random sampling of a nonmember mailing list of the Association.³ This method of distributing the questionnaire meant that any one particular company, even if very large, provided at most no more than a few respondents in the obtained sample. Thus, results cannot be attributed to the influence of the policies of any single company. In terms of types of companies, about two thirds of the sample came from manufacturing companies, while the remainder was drawn from transportation, insurance, banking, public utilities, and wholesale and retail trade firms.

The last part of the questionnaire contained a number of personal data questions, enabling the respondents to be classified on several types of independent variables. The two relevant variables used in this study were size of company for which the respondent worked and the level of his position within his organization. Classification according to the size variable was made on the basis of answers to the following question: Approximately how many employees (management and nonmanagement) are there in your company? Respondents were provided with 10 ranges of size in answering this question and were asked to check one. The range for the

smallest size was 1-49 and that for the largest was over 300,000. Arbitrarily, the sample was divided into three groups on the size dimension. The ranges and median sizes for the three categories were:

	Range of sizes	Median size
Large companies	5,000 and over	20,600
Medium companies	500-4,999	2,200
Small companies	1-499	224

From the median sizes for the three groups it can be seen that adjacent groups differed by a factor of about 10, with the two extreme groups—large and small—differing by a factor of 100.

The method of classifying respondents by the other major independent variable, vertical level of position within management, has been described in detail in a previous paper (Porter, 1962). This method resulted in the assignment of respondents to one of five level-of-position categories:

- President: Presidents and chairmen of boards
- Vice President: Vice Presidents (or their equivalents in large organizations)
- Upper-Middle: Approximately the level of division managers, plant managers, and major department managers
- Lower-Middle: Approximately the level of department and subdepartment managers
- Lower: First-level or second-level supervisors

Table 1 presents the characteristics of the sample by the three size-of-company categories: large, medium, and small. This table shows that the small company respondents were slightly younger and had slightly less formal education compared with the medium and large companies. In general, though, median age and educational level were similar among the three categories of organizations. Table 1 can also be referred to in determining the *Ns* for the various cells of respondents in future tables, where respondents are cross-classified by management level as well as by size of company.

RESULTS

Deficiencies in Need Fulfillment

Respondents' answers to Part a of each item ("How much of the characteristic is there now connected with your management position?") were subtracted from their answers to Part b of each item ("How much of the characteristic do you think should be connected with your management position?") to provide a measure of perceived deficiencies in need fulfillment. This method assumes that the larger the difference between the answers to Parts a and b of each item, the greater the perceived deficiency in fulfillment. The ra-

³ The assistance of the American Management Association, and particularly Robert F. Steadman, in obtaining the sample of respondents is gratefully acknowledged.

TABLE 1
DISTRIBUTION OF *N* OF TOTAL SAMPLE BY SIZE OF COMPANY AND LEVEL OF
MANAGEMENT, AND CHARACTERISTICS OF SAMPLE BY SIZE OF COMPANY

Size of company	Management level					Total sample	Median age	College degree (%)
	Presi- dent	Vice Presi- dent	Upper- Middle	Lower- Middle	Lower			
Large	9	136	268	248	74	735	43.8	78.2
Medium	31	280	291	149	22	773	43.2	74.4
Small	74	195	100	34	5	408	40.3	67.2

tionale for this measure, which is an indirect one derived from two direct answers of the respondents, has been explained previously (Porter, 1962).

Table 2 presents the mean differences between Parts b and a for each item for each subgroup of respondents. (Values for cells with an *N* of less than 20 have been omitted arbitrarily.) Examination of the table shows that trends in the size of deficiencies from large through medium to small companies differ depending on the level of management. Thus, there appears to be a definite interaction between management level and size of company that must be considered in drawing conclusions about the effect of company size on perceived deficiencies in need fulfillment. This interaction-type effect will be shown more clearly in the next two tables.

Table 2 also shows one other definite finding: perceived deficiencies clearly increase as one goes from higher to lower levels of management within each of the three categories of different sized companies. (The only consistent exception to this finding exists between the Upper-Middle and Lower-Middle levels in small companies, where the lower level group indicated smaller perceived deficiencies for almost all items.) Thus, conclusions reached in a preceding paper (Porter, 1962) that lower levels of management have consistently greater dissatisfaction hold up strongly even with the variable of size of company controlled. Therefore, to determine the effects of size of company on perceived deficiencies, the vertical level of positions must be taken into account; whereas, to determine the effect of level of position on

perceived deficiencies, size of company has relatively little effect because the same conclusions are reached regardless of company size.

Table 3 summarizes the trends in changes in size of mean deficiencies from large to medium to small companies. The increases and decreases in deficiencies as one goes from the large sized companies to the small sized companies are enumerated by item and management level in this table. If a mean deficiency increased by more than .05 scale units when going from large to medium companies or from medium to small companies, an increase or + was counted; whenever the mean decreased by more than .05 scale units, a decrease or - was counted. Changes of .05 or less scale units were counted as "no changes" or 0's. The last three columns under the heading of "total sample" summarize the trends for each item across all management levels. The figures presented for the total sample show quite clearly that if changes are totaled across all management levels there are no trends in any of the five need areas for smaller sized companies to have either larger or smaller perceived deficiencies in need fulfillment than larger sized companies. The reason for this lack of trend when all levels of management are combined becomes immediately apparent if the last row of the table, that for "total all items," is examined. In this last row, results are totaled across all items for each of the five management levels. Here the interaction effect of management level on size as it affects perceived deficiencies is demonstrated. If the + and - totals for each management level are ex-

TABLE 2
MEAN NEED FULFILLMENT DEFICIENCIES FOR EACH NEED CATEGORY AND ITEM:
THREE SIZES OF COMPANIES BY FIVE MANAGEMENT LEVELS

Need category	Item	Management level	Size of company		
			Large	Medium	Small
Security	I-1	President	—	0.45	0.21
		Vice President	0.36	0.40	0.58
		Upper-Middle	0.28	0.46	0.58
		Lower-Middle	0.42	0.28	0.53
		Lower	0.81	0.61	—
Social	II-1	President	—	0.39	0.29
		Vice President	0.47	0.34	0.45
		Upper-Middle	0.39	0.45	0.59
		Lower-Middle	0.40	0.44	0.29
		Lower	0.59	0.30	—
	II-2	President	—	0.03	0.48
		Vice President	0.06	0.18	0.26
		Upper-Middle	0.25	0.19	0.21
		Lower-Middle	0.26	0.21	0.26
		Lower	0.50	0.22	—
Esteem	III-1	President	—	0.45	0.35
		Vice President	0.63	0.53	0.69
		Upper-Middle	0.71	0.88	1.26
		Lower-Middle	0.87	0.99	0.91
		Lower	1.46	1.22	—
	III-2	President	—	0.35	0.17
		Vice President	0.39	0.39	0.47
		Upper-Middle	0.56	0.69	1.02
		Lower-Middle	0.82	0.97	0.79
		Lower	1.41	1.04	—
	III-3	President	—	0.35	0.16
		Vice President	0.23	0.40	0.34
		Upper-Middle	0.34	0.49	0.40
		Lower-Middle	0.35	0.35	0.26
		Lower	0.66	0.57	—
Autonomy	IV-1	President	—	0.48	0.16
		Vice President	0.56	0.67	0.69
		Upper-Middle	0.92	0.81	1.09
		Lower-Middle	1.01	0.93	1.00
		Lower	1.59	1.09	—
	IV-2	President	—	0.39	0.29
		Vice President	0.49	0.51	0.46
		Upper-Middle	0.62	0.76	0.89
		Lower-Middle	0.90	0.78	0.85
		Lower	1.30	1.00	—
	IV-3	President	—	0.32	0.28
		Vice President	0.69	0.64	0.70
		Middle-Upper	0.97	1.19	1.53
		Lower-Middle	1.24	1.41	0.97
		Lower	1.72	1.30	—

Table 2—Continued

Need category	Item	Management level	Size of company		
			Large	Medium	Small
Self-Actualization	IV-4	President	—	−0.13	−0.12
		Vice President	0.31	0.49	0.33
		Upper-Middle	0.54	0.75	1.12
		Lower-Middle	0.74	0.73	0.44
		Lower	1.12	1.13	—
	V-1	President	—	0.71	0.39
		Vice President	0.80	0.86	1.08
		Upper-Middle	0.88	1.15	1.37
		Lower-Middle	1.17	1.03	1.06
		Lower	1.51	1.57	—
	V-2	President	—	0.58	0.52
		Vice President	0.83	0.78	0.96
		Upper-Middle	1.00	1.09	1.45
		Lower-Middle	1.31	0.98	0.97
		Lower	1.53	1.09	—
	V-3	President	—	0.84	0.87
		Vice President	0.85	0.89	1.08
		Upper-Middle	1.15	1.15	1.38
		Lower-Middle	1.26	1.21	1.00
		Lower	1.53	1.43	—

amined for a significant interaction effect by a chi square test, the result is significant at the .001 level of confidence—chi square = 29.31 with 4 degrees of freedom. (It should be pointed out that such an analysis actually permits one to generalize only to other similar items for the same sample of individuals, rather than to generalize to a larger population of individuals. This is because the values for the chi square analysis are summations across different items rather than across independent samples of individuals. Nevertheless, since the present sample of managers is not only large but also drawn from a wide variety of types of companies, it is reasonable to expect that the result could be repeated for the same type of items on a new, independent sample of managers.) At the two lower levels of management, Lower and Lower-Middle, across all items deficiencies decreased from large through medium to small companies, indicating that managers in bottom levels of smaller companies were more satisfied than managers at the same levels of larger companies. At the next two levels

above, however, the findings are exactly reversed. At the Upper-Middle and Vice President levels, executives in larger companies were clearly more satisfied than their counterparts in smaller companies. At the topmost level, the President level, the picture is not entirely clear, for the figures in Table 3 for this level are based only on changes from medium to small companies. For these two categories, however, more favorable attitudes were shown in the smaller sized than in the medium sized companies.

It is also apparent from Table 3 that the trends for changes in deficiencies from larger to smaller companies within each of the five management levels, as revealed by the last row of the table, hold up quite consistently for each of the five need areas. Thus, at the Lower and Lower-Middle management levels, small company managers were more frequently satisfied (have smaller perceived deficiencies in need fulfillment) in all five need areas: Security, Social, Esteem, Autonomy, and Self-Actualization. The differences in frequencies are not large, but the direction

TABLE 3
NUMBER OF CHANGES IN SIZE OF MEAN DEFICIENCIES FROM LARGER SIZED TO
SMALLER SIZED COMPANIES WITHIN FIVE MANAGEMENT LEVELS

Need category	Item	Management level										Total sample							
		President			Vice President			Upper-Middle			Lower-Middle				Lower				
		+	0	-	+	0	-	+	0	-	+	0	-	+	0	-	+	0	-
Security	I-1	0	0	1	1	1	0	2	0	0	1	0	1	0	0	1	4	1	3
Social	II-1	0	0	1	1	0	1	2	0	0	0	1	1	0	0	1	3	1	4
	II-2	1	0	0	2	0	0	0	1	1	0	2	0	0	0	1	3	3	2
Category total		1	0	1	3	0	1	2	1	1	0	3	1	0	0	2	6	4	6
Esteem	III-1	0	0	1	1	0	1	2	0	0	1	0	1	0	0	1	4	0	4
	III-2	0	0	1	1	1	0	2	0	0	1	0	1	0	0	1	4	1	3
	III-3	0	0	1	1	0	1	1	0	1	0	1	1	0	0	1	2	1	5
Category total		0	0	3	3	1	2	5	0	1	2	1	3	0	0	3	10	2	12
Autonomy	IV-1	0	0	1	1	1	0	1	0	1	1	0	1	0	0	1	3	1	4
	IV-2	0	0	1	0	2	0	2	0	0	1	0	1	0	0	1	3	2	3
	IV-3	0	1	0	1	1	0	2	0	0	1	0	1	0	0	1	4	2	2
	IV-4	0	1	0	1	0	1	2	0	0	0	1	1	0	1	0	3	3	2
Category total		0	2	2	3	4	1	7	0	1	3	1	4	0	1	3	13	8	11
Self-Actualization	V-1	0	0	1	2	0	0	2	0	0	0	1	1	1	0	0	5	1	2
	V-2	0	0	1	1	1	0	2	0	0	0	1	1	0	0	1	3	2	3
	V-3	0	1	0	1	1	0	1	1	0	0	1	1	0	0	1	2	4	2
Category total		0	1	2	4	2	0	5	1	0	0	3	3	1	0	2	10	7	7
Total all items		1	3	9	14	8	4	21	2	3	6	8	12	1	1	11	43	22	39

TABLE 4
SUMMARY OF COMPARISONS OF SIZES OF MEAN DEFICIENCIES FOR DIFFERENT
SIZED COMPANIES WITHIN EACH MANAGEMENT LEVEL

Comparison	Frequency by management level				
	President	Vice President	Upper-Middle	Lower-Middle	Lower
Large versus small companies					
Large > Small:	—	0	0	7	—
Large = Small:	—	4	1	5	—
Large < Small:	—	9	12	1	—
Large versus medium companies					
Large > Medium:	—	2	2	5	11
Large = Medium:	—	6	1	5	1
Large < Medium:	—	5	10	3	1
Medium versus small companies					
Medium > Small:	9	2	1	7	—
Medium = Small:	3	2	1	3	—
Medium < Small:	1	9	11	3	—

TABLE 5
MEAN IMPORTANCE FOR EACH NEED CATEGORY AND ITEM:
THREE SIZES OF COMPANIES BY FIVE MANAGEMENT LEVELS

Need category	Item	Management level	Size of company		
			Large	Medium	Small
Security	I-1	President	—	5.39	5.84
		Vice President	5.45	5.49	5.37
		Upper-Middle	5.06	5.31	5.29
		Lower-Middle	5.32	5.22	5.41
		Lower	5.19	5.65	—
Social	II-1	President	—	6.26	6.08
		Vice President	6.37	6.28	6.10
		Upper-Middle	6.17	6.09	6.00
		Lower-Middle	6.05	5.91	5.59
		Lower	5.91	5.91	—
	II-2	President	—	4.32	4.61
		Vice President	4.85	4.72	4.48
		Upper-Middle	4.56	4.49	4.38
		Lower-Middle	4.73	4.74	4.29
		Lower	4.50	4.91	—
Esteem	III-1	President	—	5.06	5.15
		Vice President	5.36	5.15	5.29
		Upper-Middle	5.11	5.33	5.47
		Lower-Middle	5.24	5.26	5.21
		Lower	5.23	5.30	—
	III-2	President	—	5.71	5.48
		Vice President	5.52	5.40	5.46
		Upper-Middle	5.24	5.47	5.50
		Lower-Middle	5.42	5.54	5.21
		Lower	5.36	5.09	—
	III-3	President	—	4.94	5.43
		Vice President	5.23	5.30	5.36
		Upper-Middle	5.03	5.28	5.22
		Lower-Middle	5.03	5.14	5.18
		Lower	5.00	4.70	—
Autonomy	IV-1	President	—	6.23	6.19
		Vice President	5.88	5.84	5.84
		Upper-Middle	5.57	5.62	5.78
		Lower-Middle	5.42	5.62	5.56
		Lower	5.24	5.13	—
	IV-2	President	—	6.61	6.48
		Vice President	6.45	6.37	6.40
		Upper-Middle	6.23	6.19	6.40
		Lower-Middle	6.09	6.05	6.26
		Lower	6.16	5.96	—
	IV-3	President	—	6.61	6.48
		Vice President	6.31	6.37	6.24
		Upper-Middle	6.06	6.01	5.99
		Lower-Middle	5.86	5.89	5.53
		Lower	5.46	5.70	—

Table 5—Continued

Need category	Item	Management level	Size of company		
			Large	Medium	Small
Self-Actualization	IV-4	President	—	5.68	5.31
		Vice President	5.64	5.85	5.92
		Upper-Middle	5.55	5.68	5.92
		Lower-Middle	5.48	5.66	5.44
		Lower	5.35	5.61	—
	V-1	President	—	6.58	6.51
		Vice President	6.23	6.25	6.40
		Upper-Middle	6.23	6.32	6.41
		Lower-Middle	6.31	6.30	6.09
		Lower	6.32	6.26	—
	V-2	President	—	6.58	6.49
		Vice President	6.52	6.36	6.33
		Upper-Middle	6.27	6.27	6.38
		Lower-Middle	6.27	6.17	5.82
		Lower	6.36	6.17	—
	V-3	President	—	6.55	6.55
		Vice President	6.61	6.50	6.46
		Upper-Middle	6.46	6.38	6.44
		Lower-Middle	6.30	6.23	6.18
		Lower	6.26	6.22	—

is consistent if the figures for these two levels are combined. Similarly, if the figures for the Upper-Middle and Vice President levels are combined, managers in the larger sized companies were consistently more satisfied in all need areas than were executives in smaller companies. It therefore appears that differences in satisfaction among managers in different sized companies at each level of management are not related to particular types of needs, but rather are due to a more general feeling of satisfaction or dissatisfaction.

Table 4 summarizes the comparisons of sizes of mean deficiencies between each pair of sizes of organizations. This table shows clearly the change in trends between Lower-Middle and Upper-Middle levels of management. In each set of comparisons, large versus small, large versus medium, and medium versus small, mean deficiencies were larger in the larger sized companies below this point, and larger in the smaller sized companies above this point. (In the medium versus small comparisons, of course, another shift took place between the Vice President and

President levels, as noted previously.) This shift in trends was most clear-cut in the comparison of the two groups at the extreme ends of the size dimension, large versus small organizations.

Importance of Needs

Table 5, constructed in the same manner as Table 2, presents the data on perceived importance of the various needs for each subgroup of respondents. The values in each cell of Table 5 are based on the means of responses to Part c of each item ("How important is this to me?"). The table shows relatively few consistent trends for changes in size of mean importance from larger to smaller companies, either across all management levels or within any particular management level. Thus, in contrast to the data on perceived deficiencies, there does not seem to be any sort of systematic interaction effect of level of position on size in relation to perceived importance of needs.

The general absence of trends in Table 5 is shown more clearly in Table 6, where the

TABLE 6
NUMBER OF CHANGES IN SIZE OF MEAN IMPORTANCE FROM LARGER SIZED TO
SMALLER SIZED COMPANIES WITHIN FIVE MANAGEMENT LEVELS

Need category	Item	Management level												Total sample					
		President			Vice President			Upper-Middle			Lower-Middle						Lower		
		+	0	-	+	0	-	+	0	-	+	0	-	+	0	-	+	0	-
Security	I-1	1	0	0	0	1	1	1	1	0	1	0	1	1	0	0	4	2	2
Social	II-1	0	0	1	0	0	2	0	0	2	0	0	2	0	1	0	0	1	7*
	II-2	1	0	0	0	0	2	0	0	2	0	1	1	1	0	0	2	1	5
Category total		1	0	1	0	0	4	0	0	4	0	1	3	1	1	0	2	2	12*
Esteem	III-1	1	0	0	1	0	1	2	0	0	0	2	0	1	0	0	5	2	1
	III-2	0	0	1	1	0	1	1	1	0	1	0	1	0	0	1	3	1	4
	III-3	1	0	0	2	0	0	1	0	1	1	1	0	0	0	1	5	1	2
Category total		2	0	1	4	0	2	4	1	1	2	3	1	1	0	2	13	4	7
Autonomy	IV-1	0	1	0	0	2	0	1	1	0	1	0	1	0	0	1	2	4	2
	IV-2	0	0	1	0	1	1	1	1	0	1	1	0	0	0	1	2	3	3
	IV-3	0	0	1	1	0	1	0	2	0	0	1	1	1	0	0	2	3	3
	IV-4	0	0	1	2	0	0	2	0	0	1	0	1	1	0	0	6	0	2
Category total		0	1	3	3	3	2	4	4	0	3	2	3	2	0	2	12	10	10
Self-Actualization	V-1	0	0	1	1	1	0	2	0	0	0	1	1	0	0	1	3	2	3
	V-2	0	0	1	0	1	1	1	1	0	0	0	2	0	0	1	1	2	5
	V-3	0	1	0	0	1	1	1	0	1	0	1	1	0	1	0	1	4	3
Category total		0	1	2	1	3	2	4	1	1	0	2	4	0	1	2	5	8	11
Total all items		4	2	7	8	7	11	13	7	6	6	8	12	5	2	6	36	26	42

* $p \leq .05$.

changes in mean importance from larger to smaller sized companies are summarized. Only in the social need area was there a significant trend among the three sizes of companies. In this area, individuals in larger companies consistently regarded these needs as more important than did individuals in smaller sized companies.

DISCUSSION

Previous studies of the effects of size of work unit on job attitudes have consistently found that smaller sized units result in more favorable employee attitudes. However, these previous studies were mostly intraorganization comparisons among work groups, and arguments could be advanced for a view that interorganization comparisons might produce different results. In effect, the present study was one that made such interorganization comparisons. Therefore, the results permitted

comparisons among different sized organizations, as contrasted with the comparisons in most previous studies among different sized groups within the same organization.

The results of the present study show that there was not a clear-cut superiority of small organizations over large organizations in producing maximum job satisfaction within management. At the two lowest levels of management (and perhaps at the President level), small organization size did seem to be related to smaller perceived deficiencies in need fulfillment. However, the picture was almost exactly reversed at the Upper-Middle and Vice President levels, where managers in larger organizations indicated greater need satisfaction than those in smaller companies. Taken as a whole, the results for perceived need deficiencies do not show small companies producing more favorable attitudes across all levels of management. Apparently,

increased size of total organization may offer some advantages that are not obtained by increased size of unit within the organization. The fact that large companies usually have superior financial, technical, and human resources compared with small companies, for example, should be reflected to some degree in the need satisfaction attitudes of employees. On the other hand, larger units within an organization do not necessarily have such advantages over smaller units within the same organization, and hence increased size of intraorganizational units would not be expected to produce greater need satisfaction. In comparisons among different sized groups within the same organization only the disadvantages of increased size should affect job attitudes, whereas in comparisons among separate, different sized organizations both the disadvantages and advantages of increased size should influence job attitudes. The results of the present study perhaps offer some support of this analysis and suggest an interesting implication: Increasing the size of the total organization, and thereby achieving the technical advantages of large scale organization, will not necessarily tend to reduce the job satisfaction and morale of employees, as long as intraorganizational units are kept small.

A general proposition was advanced in the Introduction that level of position within the organizational hierarchy should have some interaction effect on the relation of organization size to job attitudes. Specifically, it was suggested that the higher the organizational level, the relatively more favorable will be the job attitudes of individuals in large organizations compared with the attitudes of those in small companies. For the four management levels up through the Vice President level, there was a definite interaction effect, in the predicted direction, of level of position on the relation between size and attitudes: Up to the Lower-Middle management level, individuals in smaller companies reported greater need satisfaction, whereas at the Upper-Middle and Vice President levels managers in larger companies indicated greater satisfaction. (The picture at the President level was not clear due to the absence of an adequate sample of presidents of large cor-

porations and the relatively small sample of presidents of middle sized firms.)

The reasons for the shift in attitudes between Lower-Middle and Upper-Middle levels in comparing large to small organizations would seem to be connected with the relative advantages and disadvantages that organization size brings to employees at different positions in the hierarchy. A good summary of the advantages of small organizations has been made by Worthy (1950):

In broader terms, the smaller organization represents a simpler social system than does the larger unit. There are fewer people, fewer levels in the organizational hierarchy, and a less minute subdivision of labor. It is easier for the employee to adapt himself to such a simpler system and to win a place in it. His work becomes more meaningful, both to him and to his associates, because he and they can readily see its relation and importance to other functions and to the organization as a whole. The organization operates primarily through the face-to-face relationships of its members and only secondarily through impersonal, institutionalized relationships. The closer relations between the individual employee and the top executive in such a situation are only one aspect—but an important one—of the relatively simple and better-integrated social system of the smaller organization [p. 173].

If the above statement is examined closely, it will be seen that the advantages of small organizations over large ones are especially relevant at the nonmanagement or worker level of organizational hierarchies, but not necessarily relevant at upper levels. Perhaps the key sentence is the one which states: "It is easier for the employee to adapt himself to such a simpler system and to win a place in it." Suppose, though, that the employee *has* adapted himself and *has won* a place in the more complex larger organization. Will he necessarily be less satisfied with need fulfillments provided by his job compared with someone who has won a similar high level position in a small organization? The results of this study suggest that there may be a point reached in the organizational hierarchy in which the advantages offered by large companies begin to outweigh the obvious disadvantages of such organizations. The dividing line would seem to come in middle management. Above this point, the greater material, financial, and personnel responsibilities, in absolute terms, of the large com-

pany manager may make him more satisfied than a comparable-level manager in the small company. Probably the broadest and firmest conclusion that can be drawn from the data on the effect of organization size on perceived deficiencies in need fulfillment is that if organization level is taken into account there is not a simple relation between size and job satisfaction within management. On the other hand, as the results also showed, size of organization has little apparent effect on the relation between organization level and satisfaction: higher level managers perceive greater need satisfaction from their jobs than do lower level managers, regardless of the size of organization for which they work.

Although size of organization, in combination with organization level, seemed to have definite effects on perceived deficiencies in need fulfillment, this variable did not have any striking effects on perceived importance of needs. In each of the five need areas, except that of social needs, managers from larger sized companies tended to attach about the same importance to a particular need area as did managers from smaller companies. In the social need area, however, individuals in the larger companies consistently considered these needs to be more important than did managers in smaller companies,

regardless of management level. This finding might indicate either that more socially oriented individuals tend to join larger rather than smaller companies, or that the size of organization has some influence on an individual's perception of the importance of social needs after he has joined a company.

REFERENCES

- MASLOW, A. H. *Motivation and personality*. New York: Harper, 1954.
- PORTER, L. W. A study of perceived need satisfactions in bottom and middle management jobs. *J. appl. Psychol.*, 1961, **45**, 1-10.
- PORTER, L. W. Job attitudes in management: I. Perceived deficiencies in need fulfillment as a function of job level. *J. appl. Psychol.*, 1962, **46**, 375-384.
- PORTER, L. W. Job attitudes in management: II. Perceived importance of needs as a function of job level. *J. appl. Psychol.*, 1963, **47**, 141-148. (a)
- PORTER, L. W. Job attitudes in management: III. Perceived deficiencies in need fulfillment as a function of line versus staff type of job. *J. appl. Psychol.*, 1963, **47**, 267-275. (b)
- STRAUSS, G., & SAYLES, L. R. *Personnel: The human problems of management*. New York: Prentice-Hall, 1960.
- VITELES, M. *Motivation and morale in industry*. New York: Norton, 1953.
- WORTHY, J. C. Organization structure and employee morale. *Amer. sociol. Rev.*, 1950, **15**, 169-179.

(Received November 26, 1962)

COMPENSATORY TRACKING WITH DIFFERENTIATING AND INTEGRATING CONTROL SYSTEMS¹

E. C. POULTON

Applied Psychology Research Unit, Cambridge, England

Tracking performance with a differentiating amplifier in the control loop, an integrating amplifier, and a simple amplifier were compared experimentally on 2 random inputs. Learning curves were obtained from separate groups of inexperienced Ss. With a high frequency input the differentiating control system gave a smaller mean error than the integrating system ($p < .001$), whereas with a medium frequency input it gave the larger mean error ($p < .02$). With both inputs the amplifying system was best ($p < .01$). After practice the differentiating system produced the smallest mean time lag but most variability, while the integrating system produced the longest time lag.

This is a sequel to an experiment on pursuit tracking (Poulton, 1963). Apart from the change in the display (Figure 1), the principal differences were: (a) that the input of lower frequency was a filtered version of the high frequency (HF) input and (b) that a more comprehensive method of oscillographic recording and scoring was used.

METHOD

Subjects. These were enlisted men in the British Royal Navy aged between 17 and 29 years. None had had much experience of tracking. Their scores on intelligence test AH4 (Heim, 1955) were found to be unrelated to their tracking performance.

Apparatus. This was the same as for the previous experiment (Poulton, 1963). A double-beam, cathode ray tube (CRT), diameter 6 inches, had a bright horizontal line $.9 \times .05$ inch fixed at its center. A spot of light, diameter .075 inch, moved in a vertical dimension and had to be held on the line. The basic input consisted in a band of white noise from about 2.5 to 40 cycles per minute, having equal energy per cycle, and an upper cutoff of about 20 decibels per octave. The HF input was an unfiltered version, while a low-pass R-C network with a time constant of .6 second was used to produce the input of medium frequency (MF).

The control was a light rod 4 inches long, which was lightly spring centered, and was held between the thumb and forefinger of the supported hand. An upward movement of the control raised the spot of light on the CRT. The control sensitivities are listed in Table 1. The error voltage was summed neglecting its sign. Simultaneous oscillographic records were also made of the input and either the response function or the movement of the control.

Experimental Design and Procedure. The number of subjects allocated to each condition in the random group design is shown in Table 1. All but one group practiced for eight sessions as indicated in Figure 2. A session consisted in five periods of tracking of just over 1.0 minute each. The procedure followed that of the previous experiment and was aimed at maximizing the rate of learning. Oscillographic recordings were made immediately after the last period of tracking in the last session.

Scoring and Calculations. The mean error with sign neglected was summed over the last 60 seconds of each period of tracking. In Figure 2 and Table 1 the mean has been expressed as a percentage of the average error which accrued over the same period when the response function was held at zero. In comparing the mean error with that in the sixth and seventh sessions of the previous experiment with a pursuit display (Poulton, 1963), the pooled data of the equivalent sessions have been used, not the data in Table 1.

In scoring the oscillographic charts, the simultaneous input and response functions were superimposed so that time and zero displacement matched (see Poulton, 1962). For each measure 10 samples were taken from each subject. Errors in amplitude were measured between the positions at which the input and corresponding response function reversed direction. Since the amplitude of the input varied from second to second, the error has been expressed

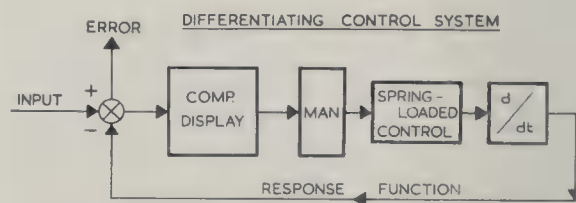


FIG. 1. The differentiating control system. (An analog integrator replaced the differentiator for the integral control task and a simple amplifier was used for the positional control task.)

¹ The subjects were supplied by the Royal Navy. Financial support from the British Medical Research Council is also gratefully acknowledged.

TABLE 1
CONDITIONS AND MEAN ERROR IN SESSIONS 7-8

Control system	Control movement for 1 inch dis- play movement	HF input			MF input		
		<i>N</i>	<i>M</i>	<i>SE</i>	<i>N</i>	<i>M</i>	<i>SE</i>
Differentiating	24° per second	6	74.1	2.1	5	56.2	1.8
Amplifying	13°	6	59.4	3.6	6	32.8	2.9
Integrating	20° for 1.0 second	5	94.7	2.3	6	45.4	2.7

Note.—All differences between means are reliable at the .02 level or better.

as a percentage of the full range of movement of the input shown on the record.

The time lag at reversals (Table 2) was the difference between when the input and corresponding response function reversed direction. Since the response functions often had no clear inflections, the time lag at inflections was the difference between the times of the inflections on the input and the times at which the response function reached the same sizes of displacement.

An "excess frequency" count was made by counting the number of reversals of the response function with an amplitude of 5% or more of the full range of movement of the input shown on the record. From this was subtracted the corresponding number for the input, and the remainder was divided by two so that it could be expressed in cycles per minute.

With the differentiating control system the time was also noted at which the input changed sign. At this instant, in the absence of a control movement, the spot of light on the face of the CRT crossed the central bright line. This time was compared with the time at which the subject reversed the direction of movement of his control stick, as was then required in order to bring the spot back to the center. With the integrating control system the time lag was measured between the reversals on the input and the reversals of the movement of the subject's control stick.

In Table 2 the standard deviations represent variability within the subjects. The standard errors of the standard deviations indicate the extent of the individual differences in variability. In comparing the differences between means, two-tailed *t* tests have always been used except where stated.

RESULTS

The learning curves are given in Figure 2. Table 1 compares the average error pooled over the seventh and eighth sessions. It shows that with both inputs the amplifying control system was the best ($p < .01$). With the HF input the integrating system was the worst ($p < .001$), while with the MF input the

differentiating system was the worst ($p < .02$). With the HF input the results for the differentiating and integrating systems are reliably worse ($p < .001$) than the comparable results with a pursuit display (Poulton, 1963, Table 1). In contrast, after practice the amplifying system was not reliably worse than with a pursuit display ($p > .05$), although it was worse on the pooled results of the first two sessions ($p < .02$).

Table 2 shows the time lags of the response functions at the end of the last session, measured at reversals and at inflections on

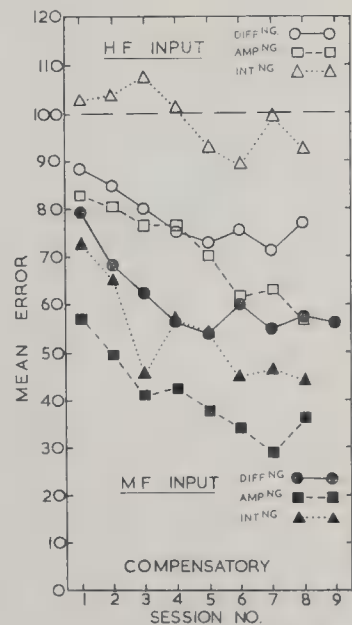


FIG. 2. Learning curves comparing the control systems and input frequencies. (Each function represents the mean performance of a group of five or six subjects. Points show the average of five 1-minute periods of tracking by each subject, adjusted to make 100 the level if the subject had not responded to all.)

TABLE 2
TIME LAG IN SECONDS AFTER PRACTICE

Control system	HF input				MF input			
	CE		SD		CE		SD	
	M	SE	M	SE	M	SE	M	SE
Time lag at reversals								
Differentiating	.07 ^b	.07	.15	.02	+.02 ^{ab}	.01	.19 ^a	.03
Amplifying	.14 ^c	.03	.14	.02	.15 ^{ac}	.01	.11 ^a	.01
Integrating	.36 ^{bc}	.08	.15	.03	.27 ^{bc}	.04	.16	.02
(control stick)	(.08)	(.08)	(.16)	(.02)	(.05)	(.02)	(.19)	(.03)
Time lag at inflections								
Differentiating	.10 ^b	.05	.20	.04	.05 ^b	.06	.25 ^{ab}	.12
Amplifying	.15 ^c	.03	.14	.03	.14 ^c	.01	.10 ^{ac}	.04
Integrating	.36 ^{bc}	.06	.18	.04	.31 ^{bc}	.05	.11 ^{bc}	.05

^a Differentiating-Amplifying $p < .02$ or better.
^b Differentiating-Integrating $p < .05$ or better.
^c Amplifying-Integrating $p < .05$ or better.

the input. The time lags of the control stick at reversals, for the integrating control system, are given in parentheses. The time lags of the response functions behave very similarly at reversals and at inflections. With both inputs the mean lag was reliably greater with the integrating system than with the differentiating, the amplifying system lying intermediately. The values in parentheses show that for the integrating system the additional time lag at reversals was produced by the 90 degrees of phase retard inserted by the control system. For the MF input the reversals of the control stick lagged in time

reliably less far behind with the integrating system than with the amplifying system ($p < .01$).

With the differentiating control system and HF input, the mean time lag between a change in the sign of the input and the corresponding change in the direction of movement of the subject's control stick which was required in order to correct it, was .08 second, $SE = .03$. This is about the same as the lag at reversals and at inflections. However the standard deviation of the error in timing here was only .09 second, $SE = .01$. This is reliably smaller than the within-subject vari-

TABLE 3
AMPLITUDE ERROR AFTER PRACTICE

Control system	HF input				MF input			
	CE		SD		CE		SD	
	M	SE	M	SE	M	SE	M	SE
Differentiating	6.1	5.0	11.1 ^a	.9	+.9	2.0	12.7 ^a	2.4
Amplifying	9.7 ^{bc}	2.8	8.7	1.1	.5 ^c	2.3	9.1	2.2
Integrating	9.1 ^{bc}	3.6	7.7 ^a	1.0	.2 ^c	2.7	7.3 ^a	1.5

Note.—The error is expressed as a percentage of the overall range of movement of the input. The mean response amplitude was too small except where indicated by a +.
^a Differentiating-Integrating $p < .05$ with a one-tailed test or better.
^b Mean reliably greater than zero at the .05 level with a one-tailed test or better.
^c HF input-MF input $p < .05$ with a one-tailed test or better.

ability at reversals and inflections ($p < .05$ or better), and suggests that zero input may have been one of the principal cues used by the subject in timing his control movements with the differentiating system. In Table 2 the differentiating system shows if anything the greatest within-subject variability in timing, although the differences are only reliable with the MF input.

Table 3 gives the amplitude errors of the response functions at the end of the last practice. The amplitude tended to be reduced with the HF input, but there was no difference in mean amplitude between the three control systems. The differentiating system showed reliably more variability from the correct amplitude than did the integrating system, the amplifying system lying intermediately.

Table 4 shows the number of cycles per minute in the response function in excess of those required to match the input. There were most with the differentiating control system and least with the integrating system. With the differentiating system they were caused by the sudden or jerky control movements made by all subjects. With the amplifying system about 40% of the excess frequency count was caused by what appeared to be incorrect anticipations of changes in the direction of movement of the input. The remainder was produced by the one or two subjects in each group who systematically superimposed up-and-down control movements of relatively small amplitude and HF upon what they reproduced of the input. This "dither" gave them a visible indication of what their control was doing, which is not normally available with a compensatory display

unless the control movements are excessively large. With the integrating system, practically the whole of the excess frequency count was produced by the one or two subjects in each group who "dithered." Here the dither came through not only the compensatory display, but also the lag introduced by the integrating system, and could thus have enabled the subject to feel more in control of the situation.

DISCUSSION

The general pattern of the results in Figure 2 and Table 1 resembles that for pursuit tracking (Poulton, 1963). As the power was increased in the higher frequencies of the input, the mean error with the integrating control system increased more rapidly and overtook the error with the differentiating system. Over the range of input frequencies tested the amplifying system was the system of choice. The differences between the amplifying and integrating control systems are in the same direction as the difference found by Chernikoff and Taylor (1957, Figure 1 and Table 2) in compensatory tracking with their fastest input.

The disadvantage of the differentiating control system was the amplification of the HFs in the subject's control movements. This is reflected directly in Table 4, and also resulted in greater variability both in timing (Table 2, MF input) and in amplitude (Table 3). The advantage of the differentiating system was the elimination of the subject's movement time; in order to change the position of the spot on the face of the CRT he had only to change the rate of movement of his control. Thus the time lags with this system were so small that they were not reliably different from zero (Table 2), whereas with the amplifying system there was always a statistically reliable lag. When the HF input was tracked with a pursuit display the differentiating system again produced a smaller time lag than the amplifying system (Poulton, 1963).

Table 2 shows that the disadvantage of the integrating control system was the time lag which it inserted. Although the subject's control movements lagged in time only

TABLE 4
EXCESS FREQUENCY COUNT AFTER PRACTICE

Control system	HF input		MF input	
	<i>M</i>	<i>SE</i>	<i>M</i>	<i>SE</i>
Differentiating	48.2	6.7	43.0	8.2
Amplifying	16.0	8.6	19.6	6.6
Integrating	5.1	5.7	11.2	7.6

Note.—Differentiating different from amplifying at the .05 level and from integrating at the .02 level or better.

about half as far behind the input as with the amplifying system, the 90 degrees of phase retard inserted by the integrating system meant that the resultant response function lagged about twice as far behind. This result is rather different from that with the HF input when a pursuit display was used (Poulton, 1963) which directly indicated the time lag in the response function. With the pursuit display, the subject actually cut the time lag of his response function with the integrating system to below that with the amplifying system by reducing its amplitude.

REFERENCES

- CHERNIKOFF, R., & TAYLOR, F. V. Effects of course frequency and aided time constant on pursuit and compensatory tracking. *J. exp. Psychol.*, 1957, **53**, 285-292.
- HEIM, A. W. *Manual for the Group Test of General Intelligence AH4*. London, England: National Foundation for Educational Research, 1955.
- POULTON, E. C. On simple methods of scoring tracking error. *Psychol. Bull.*, 1962, **59**, 320-328.
- POULTON, E. C. Pursuit tracking with differentiating and integrating control systems. *J. appl. Psychol.*, 1963, **47**, 289-292.

(Received November 27, 1962)

EVALUATION OF TYPEWRITING PROFICIENCY TRAINING: PRELIMINARY TEST DEVELOPMENT¹

LEONARD J. WEST

Southern Illinois University

AND D. J. BOLANOVICH

Perceptual Development Laboratories, St. Louis, Missouri

Development of parallel forms of a measure of office typing proficiency for use as an employment test and for assessing the results of industrial training programs used 168 persons at terminal stages of college typing training as Ss. The 2 forms do not differ in difficulty or in the shape of the distributions of scores. Reliabilities of .77 (speed) and .88 (errors) were found for the production test (business letters, tables, rough drafts), and of .97 and .73 for straight copy work. Large and significant gains in test scores followed terminal training. Validity coefficients (with dichotomized instructors' rankings) ranged between .56 and .80, with cross-validity coefficients ranging between .55 and .67 for 8, 6, 3, and 2 predictor variables. The tests appear to be satisfactory for the purposes intended.

This study reports the development of criterion tests for evaluating a typewriting training program currently being prepared by the authors that is intended primarily for use by large industrial corporations to upgrade the proficiency of experienced typists. However, the test, by its very nature, could also serve as an employment test for typists.

The requirement, then, was for alternate forms of a test that samples the major classes of skills used by industrial typists and whose forms are of equivalent difficulty, high inter-form reliability, and demonstrated validity.

Examination of pre-existing tests (Buros, 1959) revealed no commercially available materials satisfactory for these purposes. Further, Jurgensen (1942) has suggested and West (1960) has shown that the conventional measure of typewriting proficiency ("straight copy" work) has only a moderate relationship to realistic office typing tasks. Non-mechanical aspects of office typing activities, such as noting and correcting errors and

making decisions about form and arrangement, are wholly absent in straight copy work.

METHOD

Several separate but overlapping procedures were necessary to establish the information needed for development of the tests, using portions of the entire subject population, as follows:

Subjects

A total subject group of 168 different persons was used. Of these, 144 were students in advanced typing classes in the Southern Illinois University Business Education Department (53 with 1 year of previous high school typewriting training and 91 with 2 previous years of such training). These 144 persons were in six intact typing classes varying in size from 11 to 38 persons, and they were aware that their test performance would affect their grades. The remaining 24 persons were a volunteer group of experienced business teachers.

Test scores of these groups were combined for various analyses into subgroups varying from 21 to 100 subjects.

Item Construction

As inferred from an examination of typewriting textbooks and of a number of published tests, the three most prominent classes of typing activity are: correspondence, columnar or tabular material, and "rough draft" materials. Accordingly, items of these three kinds were constructed for the test. In addition, measures of straight copy proficiency were prepared in order to examine further the role of ordinary stroking skill. Two test forms were constructed, *each* consisting of:

¹ This research was supported in part by the Graduate School of Southern Illinois University and in part by Perceptual Development Laboratories. The authors wish to express their gratitude to Bonnie Lockwood for allowing us access to her classes at Southern Illinois University; to Sue Grisham for handling scoring and clerical details; and to Thomas D. Purcell, of the University's Data Processing Service, for supervising analysis of data.

TABLE 1
SPEED MEANS, STANDARD DEVIATIONS, AND DIFFERENCES BETWEEN MEANS AND
STANDARD DEVIATIONS FOR TWO TEST FORMS

Subtest	Form A		Form B		$\bar{A}-\bar{B}$	$A_{SD}-B_{SD}$
	Mean	SD	Mean	SD		
Timings ^a	59.71	10.62	60.27	9.75	-.56*	.87****
Production ^b						
Letters	44.14	12.09	43.18	12.00	.96	.09
Tables	63.59	16.94	62.08	18.56	1.51	-1.62
Drafts	60.51	15.95	60.44	15.80	.07	.15
Total production	168.24	36.96	165.70	37.30	2.54	-.34

Note.—*N* = 100.
^a In gross words per minute.
^b Completion time in $\frac{1}{4}$ minutes.
* *p* < .05.
**** *p* < .001.

1. Two unarranged business letters from typed copy: one of 100 words, the other of 130 words.
2. Two longhand tables, including columnar and main headings: one of three columns, the other of four columns—but each containing the equivalent of 50 5-stroke words.
3. Two rough drafts containing corrections to be made: one entirely in longhand and containing 80 words, the other from mixed type and longhand and containing 150 words.
4. Two pieces of ordinary prose for 3-minute straight copy timed writings, each at a syllabic intensity (mean number of syllables per word) of 1.5.

Assignment of subtest items to each form was based on preliminary administration of the items to a group of experienced business teachers at the summer 1961 session at the University of North Dakota. Mean item speed and error scores for this group confirmed the comparability of the items paired in the alternate forms.

Test Administration

Test administration was counterbalanced in each of the classes, with one random half taking the forms in A-B order, the other half in B-A order.

Scoring

Straight copy performance was scored for gross words per minute and for number of errors. The six job-type subtest items were scored for number of errors (major, minor, and total) and for completion time to the nearest $\frac{1}{4}$ minute. All errors were given unit weight.

RESULTS

The results reported bear on interform equivalence, reliability, validity, and internal task relationships. Completion time for the

six production items making up each form (straight copy excluded) ranged from 16 to 72 minutes, with a mean of 42 minutes (*SD* = 9 minutes).

Equivalence

Forms A and B were comparable in difficulty, as shown by the subtest means and standard deviations for speed (Table 1) and error (Table 2) scores.

The only significant differences were between means and between standard deviations for speed scores on straight copy timings. However, the absolute size of the differences is of little practical consequence. The two forms did not show any significant speed score differences for the production tasks.

There were no significant differences in error means or standard deviations between the two forms of the test. In addition: (a) the two forms of the test produced distributions of speed and error scores on the production tasks that did not differ significantly in shape; and (b) the speed scores were normally distributed while the error scores, as is typical, were skewed toward the high-error end.

Reliability

Correlations between Form A and Form B for total and for subtest scores are shown in Table 3.

Although some subtest intercorrelations are relatively low, the total production scores for

TABLE 2

ERROR MEANS, STANDARD DEVIATIONS, AND DIFFERENCES BETWEEN MEANS AND
STANDARD DEVIATIONS FOR TWO TEST FORMS

Subtest	Form A		Form B		$\bar{A}-\bar{B}$	$A_{SD}-B_{SD}$
	Mean	SD	Mean	SD		
Timings	12.59	7.72	13.32	7.44	-.73	.28
Production						
Letters	2.70	2.13	2.67	2.36	.03	-.23
Tables	4.81	3.92	5.02	4.06	-.21	-.14
Drafts	4.52	2.95	4.02	3.15	.50	-.20
Total production	12.03	7.28	11.71	7.79	.32	-.51

Note.— $N = 100$.

both speed and errors correlate quite highly. They are comparable to reliability coefficients for most specific ability tests in industrial use today. Timed writings typically show very high alternate-form correlations for speed scores while, in the present instance, production tasks showed higher correlations for error scores.

For the two test forms, therefore, the data show substantial equivalence both in difficulty and in the nature of the skills being measured. The obtained reliability coefficients for the forms appear to be high enough for practical use in training evaluation and for employment testing, as intended.

Validity

Two types of evidence for validity were obtained:

1. Reflection of Skill Learning: If the test is to be a valid measure of training, it should reflect known improvement in the skill being measured. Table 4 shows the pre- and post-

training differences in means for students completing terminal college typing training.

Except for a nonsignificant decrease in straight copy errors, all measures revealed significant improvement in performance (higher speeds, fewer errors). If the content of training may be accepted as appropriate for secretarial employment, then the significant and sizable performance gains following a final 12-week training period may be taken as evidence for the validity of the tests.

2. Correlation with External Criterion: A valid test of typing skill should also be expected to correlate with an external criterion of that skill. Table 5 shows point-biserial correlations between test scores (subtest and total production) and dichotomized rankings of typing skill by each of two instructors (of classes at the same terminal level of training).

The correlations between the dichotomized rankings of the two instructors and speed and error scores on the subtests and on the total production test range as shown in Table 5, with Instructor B's rankings apparently giving more weight to quality than to speed of performance. The trivial relationship between errors in straight copy work and the criterion rankings ($-.03$ and $.13$) is especially worthy of note because it has self-evident negative implications for employment tests that give appreciable weight to errors on straight copy tests—an implication further supported by the data of Table 7.

Multiple-correlation coefficients against the dichotomized rankings were computed for:

TABLE 3

ALTERNATE FORM CORRELATION COEFFICIENTS

Subtest	Speed	Errors
Timings	.97	.73
Production		
Letters	.65	.62
Tables	.42	.80
Drafts	.62	.64
Total production	.77	.88

Note.— $N = 100$.

TABLE 4
PRE- AND POSTTRAINING PERFORMANCE DIFFERENCES

Subtest	Speed means		Error means	
	Gain in speed	Increase (%)	Decrease in errors	Decrease (%)
Timings	5.24****	8.7	2.12	17.6
Production				
Letters	4.72**	9.8	1.84****	48.4
Tables	14.78****	19.5	2.31****	43.3
Drafts	7.41***	10.9	1.44**	34.6
Total production	26.91****	14.0	5.66****	42.3

Note.—Production speed in $\frac{1}{2}$ minutes; timing speed in words per minute; $N = 32$.
** $p < .02$.
*** $p < .01$.
**** $p < .001$.

all eight of the subtest variables shown in Table 5; the six production subtests (excluding straight copy performance); three variables (total production speed and major and minor errors on the production test); and two variables (total production speed and errors). The resulting R 's for each of the two instructors are displayed in the upper half of Table 6. (It may be mentioned in passing that multiple correlations for 8, 6, 3, and 2 predictor variables against a continuous-scale criterion score—available only for Instructor B—were, respectively, .80, .88, .86, and .86.)

Cross-Validation. To obtain unbiased estimates of test validity, the scoring weights of

Instructor A were used to predict the dichotomized rankings of Instructor B, and vice versa, the classes of both instructors being at the same terminal level of training. The cross-validity multiple-correlation coefficients for 8, 6, 3, and 2 predictor variables are shown in the lower half of Table 6.

For Instructor B, both originally and in cross-validation, there is little fluctuation in the validity coefficients as the number of predictors varies. Overall, and for economy's sake, the data do not appear to justify the use of more than two predictors: total production speed and total production errors. In the light of the validity (and cross-validity) coefficients of .56 and .67 (Table 6) and in view of the large and significant train-

TABLE 5
CORRELATIONS BETWEEN DICHOTOMIZED TEACHERS' RANKINGS AND PERFORMANCE SCORES

Predictor variable	Instructor A ($N = 21$)	Instructor B ($N = 32$)
Straight copy speed	.60	.40
Letter speed	.32	.19
Table speed	.48	.39
Draft speed	.32	.24
Total production speed	.47	.38
Straight copy errors	-.03	.13
Letter errors	.19	.40
Table errors	.36	.55
Draft errors	.46	.52
Total production errors	.50	.64

TABLE 6
MULTIPLE CORRELATIONS BETWEEN PREDICTORS AND CRITERION

	Number of predictor variables			
	8	6	3	2
Original validation				
Instructor A ($N = 21$)	.75	.61	.80	.56
Instructor B ($N = 32$)	.72	.70	.67	.67
Cross-validation				
A to predict B	.59	.62	.31	.67
B to predict A	.55	.55	.55	.56

TABLE 7
SUBTEST INTERCORRELATIONS FOR SUMMED FORMS

Subtest	Speed				Errors			
	Production				Production			
	Letters	Tables	Drafts	Total	Letters	Tables	Drafts	Total
Timings	-.70	-.57	-.58	-.70	.34	.22	.38	.35
Letters		.59	.70	.85		.56	.40	.72
Tables			.61	.86			.68	.93
Drafts				.89				.84

Note.— $N = 100$.

ing gains displayed in Table 4, these tests show promising validity.

Internal Task Variables

Other results of interest for both training and testing concern relationships between the skill components measured in this study.

Subtest Intercorrelations. Table 7 shows the matrix of intercorrelations of test components for both speed and error scores. The table is based on scores summed for Forms A and B. (The individual form matrices differed only very slightly from each other.) Speed-score intercorrelations (range: .58-.70) are somewhat higher than error-score intercorrelations (.22-.68). Correlations of part with total scores are, of course, inflated by overlapping; but they do give evidence of internal consistency of test components. Negative correlations of timing speed scores with production components are an artifact of the scoring measures employed (words per minute for straight copy and completion time for production components).

Although the data of Table 6 suggest little improvement in validity from part scoring, part scoring could be useful for diagnostic purposes, and there seems to be sufficient independence of components (Table 7) to

warrant it, although intercorrelations are not low.

Speed-Error Relationships. Speed-error correlations for the entire test and each of its subtests ranged between $-.08$ and $.05$, none of which is significantly different from zero. There was no tendency for the faster typists to do work of better quality. It is clearly inappropriate to evaluate a typist's skill on the basis of speed alone or of quality alone, a finding that bears on scoring. As shown in Table 3, speed and error scores did not differ appreciably in their reliabilities. In the validity studies (Tables 4, 5, and 6), however, error scores show a substantial relationship to typing skill criteria and should possibly carry higher weight than speed scores in the total criterion of performance.

REFERENCES

- BUROS, O. K. (Ed.) *Fifth mental measurements yearbook*. Highland Park, N. Y.: Gryphon Press, 1959.
- JURGENSEN, C. E. A test for selecting and training industrial typists. *Educ. psychol. Measmt.*, 1942, 2, 409-425.
- WEST, L. J. Some relationships between straight-copy typing skill and performance on job-type activities. *Delta Pi Epsilon J.*, 1960, 3(1), 17-27.

(Received November 30, 1962)

EFFECTS OF HEIGHTENED MOTIVATION ON THE DETECTION OF DECEPTION¹

LAWRENCE A. GUSTAFSON AND MARTIN T. ORNE

Massachusetts Mental Health Center and Harvard Medical School, Boston

1 of 5 cards was selected by each S and 2 minutes association to this card was required. GSR response to the selected card was compared to the responses for nonselected cards in 2 groups of Ss. 1 group was motivated to "deceive the operator and withhold responses." The other group was given no special instruction. The hypothesis that Ss who are motivated to deceive will more frequently produce disproportionately large skin resistance responses to critical items as opposed to noncritical items than will Ss who have not been so motivated was upheld. Ss who were motivated to deceive were more successfully detected. In addition detection took place at a much greater than chance level in the motivated group, while in the other group it occurred only at chance levels. The degree of autonomic response to significant stimuli appears to be a function of the motivational state of the S.

The apparent causal relationship between certain classes of verbal stimuli and physiological responses is the basis for the detection of deception by means of a polygraph. While variables which may increase or decrease the number of successful detections are often mentioned, these variables have not been manipulated experimentally.

It has been postulated repeatedly that the factor which produces the physiological response is not lying or guilt per se but rather something relating to the consequences of being detected (Burt, 1921; Chappell, 1929; Marston, 1917). This has been formulated as the punishment, or threat of punishment, theory (Davis, 1961). According to this theory, the greater the consequences of being detected, the greater the physiological response will be to the critical items, and therefore the greater the chance of detection. However, the consequences of detection have not been treated as an independent variable. This would be one of the critical tests for the theory: If no differences were found when the consequences were varied, the theory would not be valid.

On the basis of the previous observations, it

is hypothesized: Subjects who are motivated to deceive will more frequently produce disproportionately large skin resistance responses to critical items as opposed to noncritical items than will subjects who have not been so motivated. The frequency of detection among the motivated group is therefore likely to be greater.

METHOD

Subjects

Thirty-six male subjects between 18 and 25 years of age were recruited for a "paid psychological experiment" from the employment offices of four colleges in the Boston area. None of the subjects had previously participated in a study in deception. All subjects were randomly assigned to one of two experimental groups—18 to each group.

Procedure

Subjects were run individually. Upon reporting all subjects were given a very ambiguous idea of the nature of the experiment. They were told that the purpose of the study was to find out how normal subjects reacted physiologically to a series of numbers and letters, that it would be necessary to attach a number of recording electrodes to them, but that none of the electrodes would carry current to them. Electrodes for recording skin resistance (Wenger, Engle, & Clemens, 1957) and five other variables were then attached. (The remaining variables will be discussed in a later paper.) At this point a tape recording was played to half the subjects according to a random schedule, previously determined. This group was referred to as the tape group, while the remaining subjects were in the no-tape group. This recording contained the following information: (a) the experi-

¹ The research in this study was supported in part by the Institute for Experimental Psychiatry and by the Mental Health Research Training Program, Harvard Medical School.

The authors wish to express their appreciation to Emily Carota Orne for her critical comments in the preparation of this manuscript.

ment was designed to see how well the subject could keep information away from the experimenter; (b) that this was extremely difficult to do, and that only people of superior intelligence and great emotional control were able to do this; (c) they were to try as hard as they could to beat the experimenter and the equipment; and (d) if they were successful, they would be paid an extra dollar.

All subjects then picked a card from a deck of five cards. All the cards in a deck were either all number or all letter cards, with a single character on each. Half the time the letter deck was used, and half the time the number deck. The numbers were between two and nine and the letters between B and I.

The cards were arranged so the subject could not see the face of the card until he had drawn. The experimenter could not see the face at any time. After the subject had memorized the card, he placed it face down on a stand beside him. In order to make the selected card more significant to the subject, he was instructed to write down on a piece of paper, in the case of a letter card, all the words he could think of beginning with that letter, or, in the case of a number card, all the expressions and titles he could think of containing that number. He was given 2 minutes for the task. The subject was then told to lie down and relax as much as possible, and that after about 5 minutes he would hear a series of numbers (or letters), including the number (or letter) he had removed from the deck. He was not to respond verbally to any of these. The experimenter left the room and began recording the physiological measures on an Offner Type R dynograph located in an adjacent room. At the end of 5 minutes he turned on a tape recording and one item of information was presented every 15 seconds. The first item presented was a dummy foil, while the next five were the same as the characters on the five cards. After all six had been presented, they were presented in a different order. This was repeated until each character had been presented five times.

As each character was reproduced by the tape recorder, the signal pen on the polygraph was activated and the letter or number was written on the record. At the conclusion of the tape, the ex-

perimenter returned to the subjects' room and did one of three things, according to a previously arranged, randomized order. To one third of the subjects in each group he told which card they had picked, to one third he deliberately misinformed them as to which card they had picked, and to the remaining third he said nothing concerning the card they had picked. (The reasons for this design will be discussed in a separate paper.)

The subject then picked a card from a second deck. If the first deck had been numbers, the second was letters and vice versa. The remainder of the trial was exactly the same as the first trial.

The difference in skin resistance between the level immediately prior to the stimulus and the lowest level reached within 4 seconds was used as the response measure for each stimulus. Readings were made to the nearest 500 ohms. The readings were all made by a person who did not know in which group the record belonged and did not know which was the chosen letter or number.

The largest mean response was used as the predictor of the card that the subject had chosen. The mean responses for each character were determined and these means were then ranked, the largest response receiving a rank of 1. The rank of the character chosen by the subject was then determined. If this rank was 1, it was considered a correct detection, while if it was more than 1, it was considered as not successful.

RESULTS

The ranks of the selected card for Trials I and II for all subjects are shown in Table 1 for the tape and no-tape groups, along with the number of correct predictions.

A comparison was made between the ranks of the selected character for the tape and no-tape condition on Trials I and II. The Mann-Whitney *U* for Trial I was 90.0 and for Trial II was 88.5. Both of these are significant at the .05 level (two-tailed).

While primary concern of this study as put forth in the introduction was to determine

TABLE 1
MEAN RANKS FOR THE SELECTED CHARACTERS FOR DIFFERENT CONDITIONS

Trial	No-tape condition		Tape condition		Mann-Whitney <i>U</i>
	Mean rank	Number of correct predictions	Mean rank	Number of correct predictions	
I	2.36	6	1.44	12	90*
II	2.86	4	1.81	11	88.5*

Note.—Fisher exact probability tests of detectional rates for no-tape and tape conditions for Trials I and II yield: $p < .05$.
* $p < .05$, $n_1 = n_2 = 18$, two-tailed.

whether difference in subject motivation would affect the magnitude of the response to a chosen card in relationship to the magnitude of the response to other cards, it is also of interest to see how successful detection itself was in the two conditions.

Fisher exact probability tests for Trials I and II for the two conditions indicated that there was a significant difference in the number of correct detections between the tape and no-tape conditions (see Table 1). It was decided to see if detection was occurring at a greater than chance frequency in both groups. A binomial test indicated that for both Trials I and II the tape group was detected at a significantly greater than chance frequency ($p < .001$ on both trials) while for the no-tape condition this was not the case ($p > .10$ on both trials).

The records of those subjects who were in the tape group appeared to show both larger and more frequent responses, not only to the correct character but to all the characters. These differences were significant at the .05 level (Mann-Whitney U ; two-tailed).

DISCUSSION

The significant difference between the ranks of the selected character in the tape and no-tape conditions (for the first trial 1.44 and 2.36, and for the second trial 1.81 and 2.83, respectively) with corresponding differences in the relative response size to critical and noncritical items for both trials is supportive of the hypothesis put forward in the Introduction of this paper and the punishment theory of detection of deception. According to this theory the "person will give a large physiological response during lying because he anticipates serious consequences if he fails to deceive [Davis, 1961, p. 163]." In the present experiment, while the subject is paid an extra dollar if not detected, probably the greatest consequence of being detected would be a loss of self-esteem. In the tape it is mentioned that the only persons who are able to deceive are those with superior intelligence and great emotional control, two qualities which most undergraduate students cherish. (The experimenter was careful to assure the subjects who had been detected that it had been difficult and that they had put up a good

struggle.) Marston (1917) suggested that the factors which make detection possible are not directly due to the response of lying, but rather are due to an emotional reaction, probably of fear, surrounding the verbal response of lying. Burt (1921) found that having other people present during the detection procedure increased the likelihood of successful detection. In a study by Chappell (1929), it was found that simply having the subject lie without any possibility of detection or punishment did not produce any marked responses. Further, Larson (1922) had noted that after a confession, the critical items no longer produced responses. Here again the stimulus no longer produces a response after the consequences of deception have been eliminated.

Our findings by no means eliminate alternate explanations of the events underlying the detection of deception. In this experiment the consequences of being detected are quite different for the motivated group not only during the actual recording session, but also during the period when they are memorizing the card and associating to it. In terms of a conditioned response theory, it would be assumed that during the period of making associations to the card, a greater response is probably produced for the tape group than for the no-tape group because of their increased involvement and this becomes conditioned to the selected character and produces a larger response during the test situation. A future experiment which would resolve this issue would be to change the consequences of deceiving during different parts of the experiment. Certain trials could be highly rewarded for deception while other unmotivated trials could be run for an innocuous reason, such as to "check the equipment." Any differences in the relative sizes of the responses would clearly be due to differences in the consequences of being detected and not due to differences in the responses conditioned to the selected card.

The significant differences between the number of successful detections (that is the number of times a critical item was assigned a rank of 1) for the tape and no-tape conditions uphold our hypothesis that the number of successful detections is increased as

motivation is increased. It also lends support to the consequences theory mentioned earlier.

It is of interest to see what ranks were assigned to the critical items in cases where the subject was not successfully detected. In the tape condition on Trial I, of the 6 individuals who were not detected, 4 were assigned a rank of 2 on the critical number. Even in cases where detection did not occur, the critical item produced an abnormally large response and the assignment of ranks was not random. However, in the no-tape group, of the 12 who were not assigned a rank of 1 on the critical item, only 3 were assigned a rank of 2 on the critical item.

Individuals in the tape group were not able to suppress the response to the critical item, though they were able to enhance their response to one or more noncritical items.

The fact that motivated subjects were detected far more readily than chance, supports the claims made for lie detection in actual life contexts where motivation would be maximal. On the other hand, the finding that without special motivation detection in the laboratory is difficult explains some of the skepticism toward laboratory studies of deception (Berrien, 1939). Clearly the situational variables play a crucial role in the responses of the autonomic nervous system.

As mentioned in the preceding section, the tape and no-tape groups appeared to be different, not only in the number of responses made to the selected card, but to all cards. While autonomic responses are usually considered to be more or less out of the area of experimental control, except by the manipula-

tion of certain characteristics of the stimulus, such as the intensity or duration, here we find that by manipulating the role of the subject we have greatly altered his responsiveness to a stimulus which objectively remains unchanged. This relationship of the demands of the experiment to autonomic nervous system (ANS) activity is a factor that has not been considered in the discussions of response specificity and stimulus specificity. One can only speculate concerning the effect that different expectations of experimenters have on the nature of their subjects' responses. This relationship between the subject's role and ANS activity could be important for a theory of the etiology and treatment of psychosomatic disorders.

REFERENCES

- BERRIEN, F. K. A note on laboratory studies of deception. *J. exp. Psychol.*, 1939, **24**, 524-546.
- BURTT, H. E. The inspiration-expiration ratio during truth and falsehood. *J. exp. Psychol.*, 1921, **4**, 1-23.
- CHAPPELL, M. N. Blood pressure changes in deception. *Arch. Psychol.*, 1929, **17**, 5-39.
- DAVIS, R. C. Physiological responses as a means of evaluating information. In A. D. Biderman & H. Zimmer (Eds.), *The manipulation of human behavior*. New York: Wiley, 1961. Pp. 142-168.
- LARSON, J. A. The cardio-pneumo-psychogram and its use in the study of the emotions, with practical applications. *J. exp. Psychol.*, 1922, **5**, 323-328.
- MARSTON, W. M. Systolic blood pressure symptoms of deception. *J. exp. Psychol.*, 1917, **2**, 117-163.
- WENGER, M. A., ENGLE, B. T., & CLEMENS, T. K. Studies of autonomic response patterns: Rationale and methods. *Behav. Sci.*, 1957, **2**, 216-221.

(Received December 4, 1962)

DOUBLING THE RATE OF SIGNAL PRESENTATION IN A VIGILANCE TASK DURING SLEEP DEPRIVATION

D. W. J. CORCORAN

Medical Research Council, Applied Psychology Research Unit, Cambridge, England

An experiment was conducted to test the effect of doubling the amount of work required of Ss after loss of sleep. Ss were required to detect defined sequences of 3 digits ("signals") within an apparently random series. Digits were presented continuously for 30 min. either at 1 per sec. (slow) or 2 per sec. (fast). The slow condition contained 20 signals; the fast condition 40. Slow and fast groups were tested on 3 successive days of a 60-hr. vigil and 2 similar groups under control conditions. The results showed that loss of sleep affects performance under fast less than under slow. These and other results suggested that stimulation reduces the effect of sleepiness on performance.

In a study of the efficiency of teleprinter switchboard operators during a night shift (Browne, 1949), it was found that although the time operators took to answer calls rose during the early hours of the morning, the effect could be offset by an increase in the number of incoming calls. This result (having been stated) makes sense intuitively, yet there is a good *a priori* argument for predicting quite the opposite: sleepy people are best employed sleeping, that is doing nothing, so that one would expect them to be relatively better at doing next to nothing. This argument fails to take into account the reversible nature of sleepiness. It assumes that something is happening to the organism deprived of sleep, which *only sleep* can reverse. Browne's demonstration showed that increasing the demands of the work had an effect, which, although temporary, was similar to that of a period of sleep.

Such an increase in the demands of the work is not the only demonstration that certain task variables can raise the performance of sleepy subjects. Wilkinson (1961) used a form of incentive with subjects deprived of a night's sleep, and this brought their performance almost up to the levels of those who had not been kept awake. This experiment has been successfully repeated with a different kind of incentive (Corcoran, unpublished). Complexity of task (Wilkinson, 1958) and loud white noise seem to have

the same effect (Corcoran, 1962). All these variables may thus be said to act in a way similar to a period of sleep, reviving the performance of sleepy subjects, in some cases up to normal levels.

The present experiment seeks to introduce a variable rather like Browne's in a controlled laboratory setting.

METHOD

Task

The task used was a visual version of the auditory task used by Bakan (1959). The subjects lay on a couch, facing a white screen. Above and behind the subject's head was a projector, by means of which a series of apparently random digits were shown on the screen. The digits appeared singly, either at 1 per second (in the "slow" condition) or at 2 per second (in the "fast" condition). The task was continued for 30 minutes so that in the slow condition 1,800 digits appeared, and in the fast condition 3,600. Within any series of 1,800 digits, 20 sequences of "three successive odd numbers which are all different" appeared at unpredictable intervals. Subjects were requested to watch for these sequences (or "signals"), and when one was detected to call out the three numbers concerned. An equal number of signals occurred during each quarter of the test. In the slow condition the interval between signals ranged from 22 to 149 seconds; in the fast condition the intervals were halved.

Subjects and Procedure

Nineteen naval ratings between the ages of 18 and 25 served as subjects. They were divided into four groups, of which two performed during a period of 60 hours deprivation of sleep, and two served as controls. The two groups deprived of sleep, each consisting of six subjects, were treated as follows.

¹ The author wishes to thank D. E. Broadbent for his advice and the Royal Navy for supplying subjects, equipment, and assistants.

TABLE 1
PERCENTAGE OF SIGNALS DETECTED BY THE FAST AND SLOW GROUPS

	Condition	Day 1		Day 2		Day 3		N
		Mean	SD	Mean	SD	Mean	SD	
Experimental	Slow	92.9	3.71	25.8	14.75	10.0 ^a	7.07	6
	Fast	66.6	22.82	49.2	27.76	22.1	22.99	6
Control	Slow	88.8	19.53	88.8	16.77	83.2	9.61	4
	Fast	84.0	15.88	84.2	9.29	86.6	10.72	3

^a $N = 4$.

During the first week of their 6-week stay in the vicinity of the laboratory both groups were tested under the slow condition once. During the second week they slept normally on the Monday night, but did not sleep again until about 6:30 P.M. on the following Thursday. During the forenoons of Tuesday and Wednesday and the afternoon of Thursday, the subjects were again tested; that is, after approximately 4, 28, and 58 hours of wakefulness. One experimental group (the slow group) was tested on the slow condition throughout. The other experimental group (the fast group) having been tested on the slow condition during the first week was tested under fast conditions during the period of sleep deprivation. (Apparatus failure resulted in the testing of only four subjects in the experimental slow condition on the third day of their deprivation). The controls consisted of four subjects in a slow group, and three in a fast group. The testing procedure employed with these subjects was identical with that of the corresponding experimental groups except that during the intervening nights they slept normally.

RESULTS

Table 1 and Figure 1 show the average level of performance attained by the four groups on the three tests during the second week. Loss of sleep clearly had a profound effect, especially in the slow condition. The day to day drop in mean scores was significant after 1 and 2 nights without sleep. A comparison of experimentals with controls gives the following values on the Mann-Whitney U test (Siegel 1956): Day 1-2, $U = 0$, $p = .01$; Day 1-Day 3, $U = 0$, $p = .028$. The fast group did not show a significant drop between Day 1 and Day 2 ($U = 4$, $p = .262$) but did show a significant drop between Day 1 and Day 3 ($U = 0$, $p = .024$). A comparison of the percentage

drop between the slow and the fast experimentals showed that performance in the fast condition dropped less as a result of 1 and 2 nights loss of sleep ($U = 0$, $p = .002$; $U = 0$, $p = .024$, respectively). No comparable differences were evident in the control groups.

The performance changes during each of the experimental runs are shown in Figure 2. Several interesting points emerge on inspection of Figure 2. Firstly, there is no significant difference between the amount of deterioration during the run under fast and slow conditions either after 1 (Day 2) or 2 (Day 3) nights without sleep. The latter comparison is somewhat questionable, however, since no signals at all were seen after the first quarter in the slow condition on Day 3 and the data are therefore restricted on the lower level.

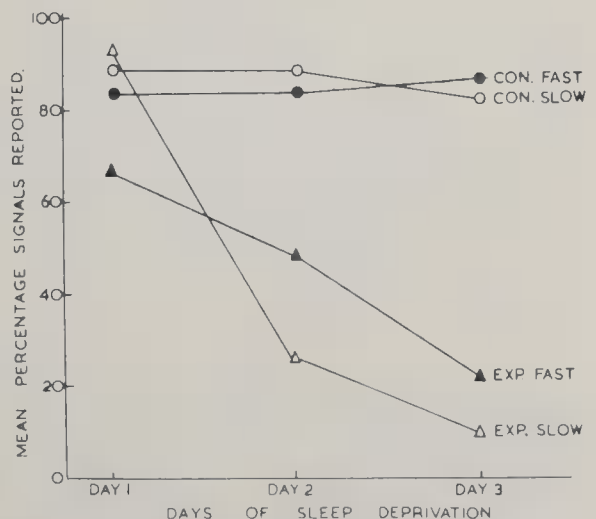


FIG. 1. Mean percentage of signals detected in 3 consecutive days.

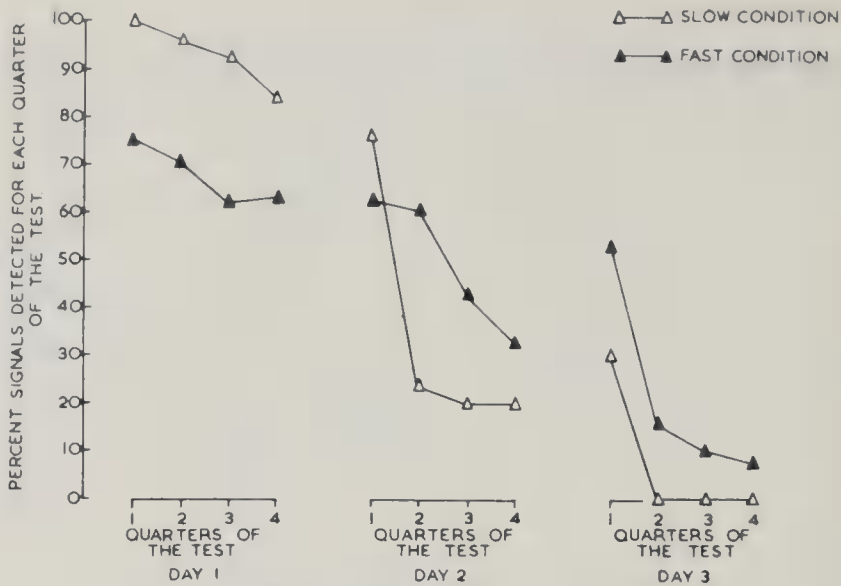


FIG. 2. Trends during the experimental runs.

Secondly, in the fast condition on Day 2 no significant decrement occurred until the third quarter of the test ($T = 3.5$, $p < .02$). On Day 3, however, a significant decrement occurred during the second quarter of the test ($T = 4$, $p < .01$). That is after 2 nights' loss of sleep on the fast condition decrement appeared earlier than it did after 1 night's loss.

Thirdly, the slow condition on Day 2 showed a decrement in the second quarter ($T = 3.84$, $p < .02$), as compared with no decrement until the *third* quarter in the fast condition. Thus it would seem likely that performing under slow conditions after 1 night's loss of sleep results in a quicker decrement in performance than performing under fast conditions. There is also an apparent similarity between the slow condition curve on Day 2 and the fast condition curve on Day 3. Altogether there is a suggestion in this data, that performing under fast conditions after 2 nights loss of sleep is rather similar to performing under slow conditions after only 1 night's loss.

DISCUSSION

These results clearly suggest that a high rate of signaling is one of the class of variables which reduce the effects of loss of sleep, and it is instructive to consider exactly what such variables have in common. While high signal frequency, high motivation, high com-

plexity and loud noise each have specific effects, each may also be said to be more "stimulating" (in rather a loose sense) than lower levels. High signal frequency, at least in the present experiment, was characterized by greater visual stimulation, greater transmission of information, and a greater response output. Complex tasks may perhaps also be said to involve the handling of greater information. High motivation may "stimulate" by creating a state of deprivation or by emphasizing the contingency of reward or punishment on efficient performance. Finally, loud noise stimulates in a very simple sense. Each variable thus stimulates, but each in rather a different way.

There is another viewpoint, which makes stimulation a reasonable explanation of the similar effects of these variables. This is the theoretical interpretation of sleepiness as lowered reticular activation (Corcoran, 1962; Wilkinson, 1958; Williams, Lubin, & Goodnow, 1959). Since the reticular system receives and responds to sensory input, it follows that increases in this input tend to determine, at least temporarily, the activation level. It is precisely this *temporary* character of the variables which distinguishes their effects from that of a period of sleep. It would therefore seem that "stimulation," in the very widest sense of the term, is effective in reducing the susceptibility of work to loss of sleep.

REFERENCES

- BAKAN, P. Extraversion-intraversion and improvement in an auditory vigilance task. *Brit. J. Psychol.*, 1959, **50**, 325-332.
- BROWNE, R. C. Day and night performance of teleprinter switchboard operators. *Occup. Psychol.*, 1949, **23**, 121-126.
- CORCORAN, D. W. J. Noise and loss of sleep. *Quart. J. exp. Psychol.*, 1962, **14**, 178-182.
- SIEGEL, S. *Nonparametric statistics for the behavioral sciences*. New York: McGraw-Hill, 1956.
- WILKINSON, R. T. The effects of lack of sleep on perception and skill. Unpublished doctoral dissertation, University of Cambridge, 1958.
- WILKINSON, R. T. Interaction of lack of sleep with knowledge of results, repeated testing, and individual differences. *J. exp. Psychol.*, 1961, **62**, 263-271.
- WILLIAMS, H. L., LUBIN, A., & GOODNOW, J. L. Impaired performance with acute sleep loss. *Psychol. Monogr.*, 1959, **73**(14, Whole No. 484).

(Received December 10, 1962)

A COMPARISON OF THREE APPROACHES TO CRITERION MEASUREMENT

FOREST O. BELL, ALVIN L. HOFF

United States Army Materiel Command, Rock Island, Illinois

AND KENNETH B. HOYT

University of Iowa

3 different criterion measures were used to assess comparative job competency of 1711 workers in 16 occupations. These were used in conjunction with the General Aptitude Test Battery (GATB) aptitude scores to establish prediction patterns of occupational success. Rating A was job oriented, Rating B was behavior oriented, and Rating C was trait oriented. The 3 scales are described and the distribution of highest r 's with GATB aptitudes is given as an empirical means of determining the relative effectiveness of the 3 approaches to criterion measurement. Conclusions are: (a) Ratings A and C are generally more useful than Rating B; (b) use of Rating A may well be justified despite the time and expense involved, if the occupations under consideration are "white-collar" and predictor aptitudes are already identified; and (c) the more economical Rating C appears to be most useful for "blue-collar" occupations and for exploratory studies to identify predictor batteries.

One of the most perplexing problems in applying psychological and aptitudinal measurements to personnel management settings is that of measuring job success for the purpose of validating employee selection tests. Few jobs are sufficiently oriented toward individual production to permit quantitative measurement. Rate of pay also presents problems in that, often, it is not directly associated with the aptitudes required for successful performance in the basic job. This is particularly true when quasi-supervisory duties are involved. Moreover, pay is frequently a function of seniority more than of job competence. As a result, most researchers have abandoned such measures in favor of some form of rating scale to estimate differing degrees of job competence. Even in this approach, there is considerable debate as to the "best" type of instrument.

Research personnel of the United States Army Ordnance Corps (which has since been integrated into the Army Materiel Command) were recently faced with this problem. Their study was designed to identify occupational prediction patterns for 16 critical civilian occupations. The primary predictor instrument used was the United States Employment Service's (USES) General Aptitude Test Battery (GATB). The battery was administered to 1,711 employees in the critical oc-

cupations. For standardization purposes, it was also given to a general Ordnance worker sample of 5,523 employees. The experiment required that the relative success of individual employees in each critical occupation be determined. For this purpose, three separate criterion measurements were used.¹ These measures were arbitrarily designated "A," "B," and "C."

Measure A was job oriented and was individually constructed by trained interviewers using a critical incident approach. They questioned the immediate supervisor of each employee regarding the job elements viewed as essential to successful performance. After these elements were listed and defined, the supervisor was asked to rate each of his employees in the sample on each element.

This process, obviously, is laborious and time consuming. In addition, it has the disadvantage of rating individual employees in an occupation against different criteria. The

¹ The results of this study are reported in detail in a report (Ordnance Civilian Personnel Field Agency, 1962) which has been supplied to most colleges and universities having a capability for offering graduate work in industrial psychology and to a substantial portion of the American Psychological Association membership registered in the Industrial Psychology Division. This paper covers certain aspects of that study which were not covered in detail in the technical report.

principal advantage is that the resulting ratings relate to specific duties of individuals. This is important since employees with the same job title frequently perform different duties. For example, one machinist may turn giant howitzer tubes while another makes tiny instrument parts requiring a precision exceeding that of most watches. Or, one personnel specialist may engage in recruitment and placement, while another operates as a salary and wage analyst.

Measure B (also completed by the worker's immediate supervisor) was oriented toward worker behavior. It consisted of 25 pairs of positive and negative behavioral descriptions. Uhrbrock (1950) suggested this approach and some of the descriptions were taken from his list. Primarily, though, descriptions were drawn from interview records collected in a previous research project and consisted of behavioral statements used frequently by Ordnance Corps supervisors in describing their employees. The descriptions were printed in random order, rather than by adjacent pairs. A score of 4 was assigned to each positive statement checked, 3 to each negative statement not checked, 2 to each positive statement not checked, and 1 to each negative statement checked.

This measure was primarily intended for use in conjunction with another instrument (the California Psychological Inventory) in a secondary objective of the study; namely, to investigate the feasibility of occupational prediction patterns based on psychological, rather than aptitudinal, factors. Nevertheless, correlation coefficients were calculated between this rating and each of the nine GATB aptitudes.

Measure C was a traditional trait oriented scale completed by second-line supervisors. Each worker was rated on 10 traits. This device was used, first, because it was employed in a majority of analogous studies known to the investigators and therefore facilitated comparison with other research findings. Second, it is easier to use than Measure A and therefore is particularly important in its implications regarding future research activity. Finally, the fact that second-line supervisors made this appraisal afforded a measure of multiple judgment as

TABLE 1

DISTRIBUTION OF CASES: HIGHEST CORRELATION OF CRITERION RATINGS WITH GATB APTITUDE SCORES

Criterion	Aptitudes								
	G	V	N	S	P	Q	K	F	M
A	8	8½	5	6	5	5	5½	8½	2
B	2	1	2	5	4	4	2	2	5
C	6	6½	9	5	7	7	8½	5½	9

Note.—All aptitudes—all occupational samples.

to the competence of each employee. This is desirable since second-line supervisors frequently have a substantial voice regarding the selection and advancement of employees.

As a part of data analysis, correlation coefficients were calculated between the mean criterion ratings and the mean aptitude scores for each of the 16 critical occupations. A comparison of the results may well be of value to other investigators in deciding what approach to take in appraising job competence.

First, it should be noted that the correlation coefficients were, in general, disappointingly low. In most instances, the "best" correlation was in the .200 to .300 range. The total range of correlations was from $-.234$ to $.369$. Negative correlations, incidentally, were not uncommon. Of the 144 correlations calculated for each of the three criterion ratings, 13 negative correlations resulted for both Ratings A and C, and 34 for Rating B.

The number of cases, in all 16 occupational samples, in which each criterion rating produced the highest correlation with each of the GATB aptitudes is indicated in Table 1. As used here, " $\frac{1}{2}$ " indicates that r 's were equal in the third decimal.²

It will be noted that Rating C produced the highest correlation with the aptitude in 62 instances, Rating A in 52 instances and Rating B in 27 instances, while Ratings A

² A 3-page table giving r 's of each criterion rating with each GATB aptitude for each occupation and correlation matrices for all criterion ratings for each occupation has been deposited with the American Documentation Institute. Order Document No. 7690 from ADI Auxiliary Publications Project, Photoduplication Service, Library of Congress, Washington, D. C. 20540. Remit in advance \$1.25 for microfilm or \$1.25 for photocopies and make checks payable to: Chief, Photoduplication Service, Library of Congress.

and C produced identical correlations with the GATB aptitudes in 3 instances. This distribution was relatively unchanged regardless of the skill level or occupational type of the job involved. (Ten of the 16 occupations can be categorized as white-collar and the remaining 6 as blue-collar occupations.)

Analysis of the data collected in the study resulted in the identification of occupational prediction patterns for 14 of the 16 critical occupations under consideration.

Occupational prediction patterns were established using an adaptation of the USES multiple cutoff procedure described by Dvorak (1956). Criterion Ratings A and C were both used in conjunction with GATB aptitude scores to arrive at the prediction patterns. When correlation with the predictor aptitudes, only, is considered, our picture changes somewhat as can be seen from Table 2. (It will be noted that only one aptitude, "F," failed to be established as a predictor in at least one occupational prediction pattern.)

The highest correlations, as indicated in Table 2, are as follows: with Rating A, 22 instances; with Rating B, 5 instances; with Rating C, 14 instances; while correlations with Ratings A and C are equal in 1 instance. This pattern is even more pronounced when correlations between the criterion ratings and predictor aptitudes for the white-collar occupations, only, are considered. Highest correlations then are: with Rating A, 15; with Rating B, 4; and with Rating C, 7. (This, of course, means that when blue-collar patterns only are considered, Ratings A and C each produced the highest correlation with predictor aptitude scores in 7 instances and were equal in one sample, while Rating B results in the highest correlation in 1 instance.)

These findings are far from conclusive.

TABLE 2
DISTRIBUTION OF CASES: HIGHEST CORRELATION OF
CRITERION RATINGS WITH GATB APTITUDE SCORES

Criterion	Aptitudes								
	G	V	N	S	P	Q	K	F	M
A	6	2	4	2	2	4	$\frac{1}{2}$	0	1
B	1	0	1	0	1	1	0	0	0
C	5	0	3	1	2	3	$\frac{1}{2}$	0	2

Note.—Predictor aptitudes only—all occupations for which prediction patterns resulted.

There are, however, some implications which may well warrant consideration in choosing a criterion measure for other studies of this type. First, while there are believed to be distinct advantages in the use of multiple criterion ratings, this procedure is recognized as being both time consuming and expensive. The question then becomes, if only one criterion rating is to be used, what sort of device is likely to yield the best results?

To this question, there would appear to be two answers. The trait oriented rating would appear to be the logical choice under some circumstances. For example, if the study involves the use of a sizeable battery from which predictors are to be identified, the efficiency of the trait oriented measure appears likely to be as great or greater than that of the other types, especially if the jobs involved are of the blue-collar variety. This is particularly true when economy and ease of administration are considered.

On the other hand, if the predictors have already been identified, *and* white-collar jobs are involved in the study, the job oriented instrument such as Measure A, described above, would appear to have a considerable advantage. Even recognizing the time, effort, and expense involved, such a decision could well be entirely justifiable.

One final note in this regard is that the correlation coefficients between Ratings A and C were, with one exception, in the .430 to .698 range. The single exception yielded a correlation of .236. Correlations between Ratings A and B were in the range of .286 to .616; and between Ratings B and C, the range was .177 to .605. Except for the single case noted above, distribution within each range was fairly even. Thus, it is obvious, that none of the three ratings can be considered to be highly comparable with any other.

REFERENCES

DVORAK, BEATRICE J. Advantages of the multiple cut-off method. *Personnel Psychol.*, 1956, 9, 45-47.
ORDNANCE CIVILIAN PERSONNEL FIELD AGENCY. Selecting employees for developmental opportunities. Technical report, July 1962, United States Army Ordnance Corps, Rock Island, Illinois.
UHRBROCK, R. S. Standardization of 724 rating scale statements. *Personnel Psychol.*, 1950, 3, 285-316.

(Received December 12, 1962)

Journal of Applied Psychology

KENNETH E. CLARK, Editor
University of Rochester

Consulting Editors

GEORGE E. BRIGGS, *Ohio State University*

MARVIN D. DUNNETTE, *University of Minnesota*

NORMAN FREDERIKSEN, *Educational Testing Service*

LEONARD D. GOODSTEIN, *University of Cincinnati*

EDWIN R. HENRY, *Standard Oil Company of New Jersey*

JOHN HOLLAND, *American College Testing Program*

CLIFFORD E. JURGENSEN, *Minneapolis Gas Company*

LAURENCE S. MCGAUGHRAN, *University of Houston*

QUINN MCNEMAR, *Stanford University*

HAROLD F. ROTHE, *Beloit Corporation*

THOMAS A. RYAN, *Cornell University*

ALEXANDER G. WESMAN, *Psychological Corporation*

CLARK L. WILSON, *Harvard Business School*

Volume 48, 1964

HELEN ORR
Managing Editor

ELIZABETH S. REED
Advertising Manager

FRANCES L. BREWER
Editorial Assistant

Published bimonthly by the American Psychological Association, Inc.
Prince and Lemon Sts., Lancaster, Pa. 17604 and 1200 Seventeenth St. N.W.
Washington, D. C. 20036

Second-class postage paid at Lancaster, Pa.

© 1964 by the American Psychological Association, Inc.

Journal of Applied Psychology

KENNETH E. CLARK, Editor
University of Rochester

Consulting Editors

GEORGE E. BRIGGS, *Ohio State University*

MARVIN D. DUNNETTE, *University of Minnesota*

NORMAN FREDERIKSEN, *Educational Testing Service*

LEONARD D. GOODSTEIN, *University of Cincinnati*

EDWIN R. HENRY, *Standard Oil Company of New Jersey*

JOHN HOLLAND, *American College Testing Program*

CLIFFORD E. JURGENSEN, *Minneapolis Gas Company*

LAURENCE S. MCGAUGHRAN, *University of Houston*

QUINN MCNEMAR, *Stanford University*

HAROLD F. ROTHE, *Beloit Corporation*

THOMAS A. RYAN, *Cornell University*

ALEXANDER G. WESMAN, *Psychological Corporation*

CLARK L. WILSON, *Harvard Business School*

Volume 48, 1964

HELEN ORR
Managing Editor

ELIZABETH S. REED
Advertising Manager

FRANCES L. BREWER
Editorial Assistant

Published bimonthly by the American Psychological Association, Inc.
Prince and Lemon Sts., Lancaster, Pa. 17604 and 1200 Seventeenth St. N.W.
Washington, D. C. 20036

Second-class postage paid at Lancaster, Pa.

© 1964 by the American Psychological Association, Inc.

Contents of Volume 48

Albrecht, Paul A., Glaser, Edward M., and Marks, John. Validation of a Multiple-Assessment Procedure for Managerial Personnel	351
Allen, Bernadene V. See Matarazzo, Joseph D.	
Anderson, Lynn R., and Fiedler, Fred E. The Effect of Participatory and Supervisory Leadership on Group Creativity	227
Andersson, Bengt-Erik, and Nilsson, Stig-Göran. Studies in the Reliability and Validity of the Critical Incident Technique	398
Astin, Alexander W., and Nichols, Robert C. Life Goals and Vocational Choice	50
Bachman, Jerald G. Prediction of Academic Achievement Using the Edwards Need Achievement Scale	16
Baldwin, R. D., Wright, A. D., and Lehr, D. J. Relation between Radar Detection and the Observer's Concept of a Target	81
Barnette, Jr., W. Leslie, and McCall, John N. Validation of the Minnesota Vocational Interest Inventory for Vocational High School Boys	378
Barrett, Gerald. See Svetlik, Byron.	
Barrett, Richard S. See Goldberg, Myles H.	
Bedrosian, Hrach. An Analysis of Vocational Interests at Two Levels of Management	325
Berdie, Ralph F. See Strong, E. K. Jr.	
Bourdon, Roger D. See Madden, Joseph M.	
Breitrose, Henry S. See Jecker, Jon.	
Brenner, Marshall H. Test Difficulty, Reliability, and Discrimination as Functions of Item Difficulty Order	98
Brilhart, John K., and Jochem, Lurene M. Effects of Different Patterns of Outcomes of Problem-Solving Discussion	175
Buel, William D. Voluntary Female Clerical Turnover: The Concurrent and Predictive Validity of a Weighted Application Blank	180
Cambria, Richard. See Smith, Karl U.	
Campbell, David P. See Strong, E. K., Jr.	
Chaney, Frederick B., and Owens, William A. Life History Antecedents of Sales, Research, and General Engineering Interest	101
Chapanis, Alphonse. Knowledge of Performance as an Incentive in Repetitive, Monotonous Tasks	263
Clark, Kenneth E. See Strong, E. K. Jr.	
Coleman, E. B. The Comprehensibility of Several Grammatical Transformations	186
Corcoran, D. W. J. The Influence of Task Complexity and Practice on Performance after Loss of Sleep	339
Cox, John A. Application of a Method of Evaluating Training	84
Darley, John G. Edward Kellogg Strong, Jr.: 1884-1963	73
Dashevsky, Sidney G. Check-Reading Accuracy as a Function of Pointer Alignment, Patterning, and Viewing Angle	344
Dashevsky, Sidney G. Combining Check-Reading Accuracy and Quantitative Information in a Space-Saving Display	348
Datta, Lois-Ellin. A Note on the Remote Associates Test, United States Culture, and Creativity	184
Datta, Lois-Ellin. Remote Associates Test as a Predictor of Creativity in Engineers	183
Dawson, Robert I. See Goldberg, Myles H.	
Dudek, Frank J., and Thoman, Evelyn. Scaling Preferences for Television Shows	237
Ehrle, Raymond A. Quantification of Biographical Data for Predicting Vocational Rehabilitation Success	171
Ewen, Robert B. Some Determinants of Job Satisfaction: A Study of the Generality of Herzberg's Theory	161
Fiedler, Fred E. See Anderson, Lynn R.	
Friedlander, Frank. Job Characteristics as Satisfiers and Dissatisfiers	388
Glaser, Edward M. See Albrecht, Paul A.	
Goldberg, Myles H., Dawson, Robert I., and Barrett, Richard S. Comparison of Programmed and Conventional Instruction Methods	110
Gough, Harrison G. A Cross-Cultural Study of Achievement Motivation	191
Gough, Harrison G., and Hall, Wallace B. Prediction of Performance in Medical School from the California Psychological Inventory	218
Gould, John D. Stereoscopic Television Pursuit Tracking	369
Gould, John D., and Smith, Karl U. Sensory-Feedback Analysis of Stereotelevision Pursuit Tracking	152
Gould, John D. See Smith, Karl U.	
Guest, Lester. Brand Loyalty Revisited: A Twenty-Year Report	93

Guion, Robert M., and Robins, James E. A Note on the Nagle Attitude Scale	29
Gustafson, Lawrence A., and Orne, Martin T. The Effects of Task and Method of Stimulus Presentation of the Detection of Deception	383
Hall, Wallace B. See Gough, Harrison G.	
Harrell, Thomas W. See Williams, Frank J.	
Henry, Mildred M. See Porter, Lyman W.	
Hill, A. B., and Large, C. Q. The Effects of Time Stress and the Elimination of Cue Information on the Display-Control Relationships of Moving Scale Instruments	255
Hinrichs, John R. The Attitudes of Research Chemists	287
Hudson, Wellborn R. See Simon, J. Richard.	
Hulin, Charles L., and Smith, Patricia C. Sex Differences in Job Satisfaction	88
Hulin, Charles L. See Locke, Edwin A.	
Hunt, Raymond G. See McMahon, Frank B. Jr.	
Jacobsen, Tony L. See Taylor, Calvin, W.	
Jecker, Jon, Maccoby, Nathan, Breitrose, Henry S., and Rose, Ernest D. Teacher Accuracy in Assessing Cognitive Visual Feedback from Students	393
Jeffrey, Thomas E. See Jones, Lyle V.	
Jenkins, William S. See Wiley, Llewellyn.	
Jerdee, Thomas H. Supervisor Perception of Work Group Morale	259
Jochem, Lurene M. See Brillhart, John K.	
Jones, Francis E. Predictor Variables for Creativity in Industrial Science	134
Jones, Lyle V., and Jeffrey, Thomas E. A Quantitative Analysis of Expressed Preferences for Compensation Plans	201
Kendall, Lorne M. See Locke, Edwin A.	
Kershner, Alan M. Speed of Reading in an Adult Population under Differential Conditions	25
Kidd, J. S., and Micocci, Angelo. Maintenance of Vigilance in an Auditory Monitoring Task	13
Knapp, Robert R. Value and Personality Differences between Offenders and Nonoffenders	59
Large, C. Q. See Hill, A. B.	
Lehr, D. J. See Baldwin, R. D.	
Locke, Edwin A., Smith, Patricia C., Kendall, Lorne M., Hulin, Charles L., and Miller, Anne M. Convergent and Discriminant Validity for Areas and Methods of Rating Job Satisfaction	313
McCall, John N. See Barnette, Jr., W. Leslie.	
McDermid, Charles, and Smith, Karl U. Compensatory Reaction to Angularly Displaced Visual Feedback in Behavior	63
McMahon, Frank B. Jr., and Hunt, Raymond G. A "Contingent-Item" Method for Constructing a Short Personality Questionnaire	197
Maccoby, Nathan. See Jecker, Jon.	
Madden, Joseph M., and Bourdon, Roger D. Effects of Variations in Rating Scale Format on Judgment	147
Marks, John. See Albrecht, Paul A.	
Matarazzo, Joseph D., Allen, Bernadene V., Saslow, George, and Wiens, Arthur. Characteristics of Successful Policemen and Firemen Applicants	123
Micocci, Angelo. See Kidd, J. S.	
Miller, Anne M. See Locke, Edwin, A.	
Moss, Stanley M. Tracking with a Differential Brightness Display: I. Acquisition and Transfer	115
Moss, Stanley M. Tracking with a Differential Brightness Display: II. Peripheral Tracking	249
Myers, James H. A Factor Analysis of Retail Credit Application Data	168
Naylor, James C. Accuracy and Variability of Information Sources as Determiners of Performance and Source Preference of Decision Makers	43
Naylor, J. C. See Rizzo, John R.	
Nealey, Stanley M. Determining Worker Preferences among Employee Benefit Programs	7
Nelson, Paul D. Supervisor Esteem and Personnel Evaluations	106
Nichols, Robert C. See Astin, Alexander W.	
Nilsson, Stig-Göran. See Andersson, Bengt-Erik.	
Orne, Martin T. See Gustafson, Lawrence A.	
Owens, William A. See Chaney, Frederick B.	
Parker, Howard A. See Taylor, James B.	
Porter, Lyman W., and Henry, Mildred M. Job Attitudes in Management: V. Perceptions of the Importance of Certain Personality Traits as a Function of Job Level	31
Porter, Lyman W., and Henry, Mildred M. Job Attitudes in Management: VI. Perceptions of the Importance of Certain Personality Traits as a Function of Line versus Staff Type of Job	305
Price, Philip B. See Taylor, Calvin W.	
Prien, Erich. See Svetlik, Byron.	
Richards Jr., James M. See Taylor, Calvin W.	

Rizzo, John R., and Naylor, J. C. The Factorial Structure of Selected Consumer Choice Parameters and Their Relationship to Personal Values	241
Roadman, Harry E. An Industrial Use of Peer Ratings	211
Robins, James E. See Guion, Robert M.	
Rose, Ernest D. See Jecker, Jon.	
Saleh, Shoukry D. A Study of Attitude Change in the Preretirement Period	310
Saslow, George. See Matarazzo, Joseph D.	
Schultz, Douglas G., and Siegel, Arthur I. The Analysis of Job Performance by Multidimensional Scaling Techniques	329
Seltzer, Carl C. Occupation and Smoking in College Graduates	1
Siegel, Arthur I. See Schultz, Douglas G.	
Simon, J. Richard. Magnification as a Variable in Subminiature Work	20
Simon, J. Richard, and Hudson, Wellborn R. An Experimental Study of the Relation between Nursing Care and Patient Welfare	268
Smith, Karl U., Cambria, Richard, and Steffan, James. Sensory-Feedback Analysis of Reading	275
Smith, Karl U., and Gould, John D. Sensory-Feedback Analysis of Behavior in Stereotelevised Visual Fields	361
Smith, Karl U. See Gould, John D.	
Smith, Karl U. See McDermid, Charles.	
Smith, Patricia Cain. See Hulin, Charles L.	
Smith, Patricia Cain. See Locke, Edwin A.	
Smith, Sidney L., and Thomas, Donald W. Color versus Shape Coding in Information Displays	137
Steffan, James. See Smith, Karl U.	
Steinemann, John H. Use of a Logically Related Predictor in Determining Intragroup Differential Predictability	336
Strong, E. K., Jr., Campbell, David P., Berdie, Ralph F., and Clark, Kenneth E. Proposed Scoring Changes for the Strong Vocational Interest Blank	75
Svetlik, Byron, Prien, Erich, and Barrett, Gerald. Relationships between Job Difficulty, Employee's Attitude toward his Job, and Supervisory Ratings of the Employee Effectiveness	320
Taylor, Calvin W., Price, Philip B. Richards, Jr., James M., and Jacobsen, Tony L. An Investigation of the Criterion Problem for a Medical School Faculty	294
Taylor, James B., and Parker, Howard A. Graphic Ratings and Attitude Measurement: A Comparison of Research Tactics	37
Thoman, Evelyn. See Dudek, Frank J.	
Thomas, Donald W. See Smith, Sidney L.	
Whitehill, Arthur M. Jr. Cultural Values and Employee Attitudes: United States and Japan	69
Wiens, Arthur. See Matarazzo, Joseph D.	
Wiley, Llewellyn, and Jenkins, William S. Selecting Competent Raters	215
Williams, Frank J., and Harrell, Thomas W. Predicting Success in Business	164
Willingham, Warren W. Estimating the Number of Different Selection Decisions Resulting from the Use of Alternate Predictor Composites	302
Wright, A. D. See Baldwin, R. D.	

DETROIT, MICHIGAN
PLEASE DO NOT REMOVE

Journal of Applied Psychology

KENNETH E. CLARK, Editor
University of Rochester

Table of Contents

Occupation and Smoking in College Graduates.....	Carl C. Seltzer	1
Determining Worker Preferences among Employee Benefit Programs.....	Stanley M. Nealey	7
Maintenance of Vigilance in an Auditory Monitoring Task....	J. S. Kidd and Angelo Micocci	13
Prediction of Academic Achievement Using the Edwards Need Achievement Scale.....	Jerald G. Bachman	16
Magnification as a Variable in Subminiature Work.....	J. Richard Simon	20
Speed of Reading in an Adult Population under Differential Conditions....	Alan M. Kershner	25
A Note on the Nagle Attitude Scale.....	Robert M. Guion and James E. Robins	29
Job Attitudes in Management: V. Perceptions of the Importance of Certain Personality Traits as a Function of Job Level.....	Lyman W. Porter and Mildred M. Henry	31
Graphic Ratings and Attitude Measurement: A Comparison of Research Tactics.....	James B. Taylor and Howard A. Parker	37
Accuracy and Variability of Information Sources as Determiners of Performance and Source Preference of Decision Makers.....	James C. Naylor	43
Life Goals and Vocational Choice.....	Alexander W. Astin and Robert C. Nichols	50
Value and Personality Differences between Offenders and Nonoffenders....	Robert R. Knapp	59
Compensatory Reaction to Angularly Displaced Visual Feedback in Behavior.....	Charles McDermid and Karl U. Smith	63
Cultural Values and Employee Attitudes: United States and Japan..	Arthur M. Whitehill, Jr.	69

American Psychological Association

Consulting Editors

GEORGE E. BRIGGS, *Ohio State University*
NORMAN FREDERIKSEN, *Educational Testing Service*
LEONARD D. GOODSTEIN, *University of Iowa*
EDWIN R. HENRY, *Peace Corps*
JOHN HOLLAND, *American College Testing Program*
CLIFFORD E. JURGENSEN, *Minneapolis Gas Company*

LAURENCE S. MCGAUGHRAN, *University of Houston*
QUINN MCNEMAR, *Stanford University*
HAROLD F. ROTHE, *Beloit Corporation*
THOMAS A. RYAN, *Cornell University*
ALEXANDER G. WESMAN, *Psychological Corporation*
CLARK L. WILSON, *Batten, Barton, Durstine, and Osborn, Incorporated*

This journal gives primary consideration to original investigations in any field of applied psychology except clinical psychology, although a descriptive or theoretical article may be accepted if it represents a special contribution in an applied field. Quantitative investigations of interest or value to psychologists working in the following broad fields will be considered: vocational and educational prognosis, diagnosis, and guidance at the secondary and college level; personnel research in business, industry, and government; engineering psychology; industrial working conditions; research on opinion and morale factors; job analysis and classification research; market and advertising research.

Manuscripts must be accompanied by an abstract of 100–120 words typed on a separate sheet of paper. The abstract should conform to the style of *Psychological Abstracts*. Detailed instructions for preparation of the abstracts appeared in the *American Psychologist* (1961, 16, 833), or they may be obtained from the Editor or from the APA Central Office.

Manuscripts should be addressed to the Editor:

Dr. Kenneth E. Clark
College of Arts and Science
302 Morey Hall
University of Rochester
Rochester, N. Y. 14627

All manuscripts must be submitted in duplicate. Original figures are prepared for publication; duplicate figures may be photographs or pencil-drawn copies.

Manuscripts must conform to the style requirements described in the *Publication Manual of the American Psychological Association*.

The following policies govern reprinting of materials copyrighted in APA journals: (a) to require approval to reprint for tables and figures, and for text only if more than 500 words [total from one article] in length; (b) to continue approval to reprint, contingent on the author's approval, for articles reprinted in whole or in major part; (c) to negotiate where possible the dedication of royalties from commercial publishers to the American Psychological Foundation.

Journal of Applied Psychology

Published bimonthly by the
American Psychological Association
Prince and Lemon Sts., Lancaster, Pa. 17604
and 1333 Sixteenth St. N. W.
Washington, D. C. 20036

\$10.00 per volume

\$2.00 per issue

HELEN ORR
Managing Editor

ELIZABETH S. REED
Advertising Manager

VIRGINIA RICHARDS
Editorial Assistant

Subscriptions, orders, and business communications should be addressed to the American Psychological Association, 1333 Sixteenth St. N. W., Washington, D. C. 20036. Address changes must reach the subscription office by the tenth of the month to take effect the following month. Undelivered copies resulting from address changes will not be replaced; subscribers should notify the post office that they will guarantee second-class forwarding postage. Other claims for undelivered copies must be made within four months of publication.

Second class postage paid at Lancaster, Pa. and at additional mailing offices.

© 1964 by the American Psychological Association, Inc.

Journal of Applied Psychology

VOL. 48, No. 1

FEBRUARY 1964

OCCUPATION AND SMOKING IN COLLEGE GRADUATES¹

CARL C. SELTZER²

The results of a study of 895 members of the Harvard Class of 1946, 13 years after graduation, with respect to the association of occupation and smoking behavior, indicate statistically significant differentiations between smokers and nonsmokers; between cigarette, cigar, and pipe smokers; and in accordance with degree or rate of cigarette smoking. The significance of these findings appears to relate to the influence of personality and constitution on smoking behavior.

With the problem of smoking and health continuing to require more intensive investigation, increased attention is being focused on the determination of the host factors involved in tobacco smoking, the nature and behavior of the persons who participate in this form of activity. From the research done thus far, it has become increasingly clear that the smoking habit, the form of smoking adopted, and the abstention therefrom, are not simply superficial, indiscriminate, fortuitous behavioral patterns resulting from sociocultural influences in society (Heath, 1958; Matarazzo & Saslow, 1960; Seltzer, 1963). On the contrary, the evidence suggests that they are structured reflections of very complex forces, innate and environmental in constant counterplay. Smoking behavior appears to be a response to a wide variety of personality and behavioral characteristics which, in part, have their origin in the biological or genetic makeup of the individual, this predisposition leading men to "seek different environments and to react to them in different ways." Concomitantly, the nature of the environmental opportunities—social, cultural, and economic—

is a vital element in conditioning the development and expression of the individual's behavior pattern.

It is in this context that studies of occupational patterns of smokers and nonsmokers take on significance, in providing additional information relative to the host factors involved in smoking behavior.

The present article is concerned with the occupational patterns of different classes of nonsmokers, cigarette smokers, pipe smokers, and cigar smokers, in a group of male college graduates from a large eastern university. The analysis deals with a comparison of smokers and nonsmokers, variations in occupational patterns according to form of smoking adopted, and variations in occupational patterns as related to degree or rate of cigarette smoking.

METHOD

The data upon which this study is based were derived from 895 members of the Harvard Class of 1946 as part of an investigation of morphological constitution and smoking (Seltzer, 1963). Information on occupation and smoking habits were obtained by questionnaire mailed to 1,138 members of the Class, resulting in a response by 922 or 81% of the group. In 27 instances no occupational information was present in the questionnaire, thereby reducing the number to a basic series of 895 individuals.

Since at the time of the reply to the questionnaire these Harvard graduates were already 13 years out of college and averaged 35 years of age,

¹ This study was supported by a grant from the Tobacco Industry Research Committee.

² Research Fellow in Anthropology, Peabody Museum, Harvard University and Research Associate in Physical Anthropology, Department of Nutrition, Harvard School of Public Health, Boston, Massachusetts.

the smoking histories covered a considerable period and their occupational status was for the most part firmly established. Current smokers in the group had been engaged in tobacco smoking on the average for somewhat over 15 years, while the discontinued or ex-smokers gave a history of more than 11 years.

Smoking categories were established with a view of obtaining groupings as precisely differentiated as possible. The primary classification divided the subjects into nonsmokers and smokers. The nonsmoker was defined as a person who had never smoked at all or had only attempted an occasional smoke during his lifetime. Individuals who smoked occasionally but not every day were excluded from the nonsmoker category. The smokers in the series were subdivided into exclusive groupings of "pure" cigarette, "pure" pipe, and "pure" cigar smokers in accordance with the form of tobacco used. Accordingly, all "mixed" smokers, those who regularly used more than one form of tobacco, were

omitted from this particular classification. For the analysis of the rate of cigarette smoking, the category of cigarette smokers was assorted into four subgroups based on daily rate of consumption as follows: less than 10 cigarettes daily, $\frac{1}{2}$ -1 pack daily (10-20 cigarettes), 1-2 packs daily (20-40 cigarettes), 2 + packs daily (40 + cigarettes). In this instance no distinction was made as to pure or mixed smokers.

Considerable attention was given to the adoption of an occupational classification for this study. An extensive number of classifications were available. However, the highly selective nature of our subjects and the special orientation of this study toward obtaining an understanding of the "self-selective" and personality factors of the individuals with the smoking habit were important elements affecting the choice of classification. The system developed by Roe (1956) was found to be most suitable for the material and purpose on hand. In this classification, group categorization is by

TABLE 1
OCCUPATION GROUPS: COMPARISON BETWEEN SMOKERS AND NONSMOKERS

Group	Nonsmokers		Smokers		χ^2
	No.	(%)	No.	(%)	
Business Contact	15	6.7	77	11.5	4.25*
Organization					
Major and minor executives	50	22.2	199	29.7	4.70*
Bankers	6	2.7	21	3.1	0.12
Total	56	24.9	220	32.8	4.99*
Technology	23	10.2	46	6.9	2.66
Outdoor	3	1.3	7	1.0	0.13
Science					
Research scientists	14	6.2	44	6.6	0.03
Physicians	22	9.8	72	10.7	0.16
Surgeons	9	4.0	12	1.8	3.58
Total	45	20.0	128	19.1	0.08
General Cultural					
Cultural administrators	3	1.3	20	3.0	1.84
Attorneys	31	13.8	75	11.2	1.09
College professors	12	5.3	29	4.3	0.38
Clergymen	6	2.7	9	1.3	1.79
Teachers (high and elementary)	10	4.4	19	2.8	1.39
Others (journalists, librarians)	14	6.2	16	2.4	7.64**
Total	76	33.7	168	25.0	6.43**
Arts and Entertainment	7	3.1	24	3.6	0.11
Grand total	225	100.0	670	100.0	

Note.—Chi squares based on a 2 × 2 classification. The overall association of the whole table yields a $\chi^2 = 26.87$ and $p < .02$.
* $p < .05$.
** $p < .01$.

TABLE 2
OCCUPATION GROUPS: COMPARISON BETWEEN PURE CIGARETTE,
PURE PIPE, AND PURE CIGAR SMOKERS

Group	Pure cigarette		Pure pipe		Pure cigar		χ^2
	No.	(%)	No.	(%)	No.	(%)	
Business Contact	62	14.3	8	13.6	1	3.0	3.33
Organization							
Major and minor executives	134	30.9	8	13.6	14	42.4	9.65**
Bankers	12	2.8	2	3.4	0	0.0	0.03
Total	146	33.7	10	16.9	14	42.4	8.78*
Technology	29	6.7	4	6.8	2	6.1	0.02
Outdoor	3	0.7	1	1.7	1	3.0	1.87
Science							
Research scientists	22	5.1	6	10.2	3	9.1	3.06
Physicians	48	11.1	4	6.8	2	6.1	1.72
Surgeons	11	2.5	0	0.0	0	0.0	2.40
Total	81	18.7	10	16.9	5	15.2	0.33
General Cultural							
Cultural administrators	11	2.5	1	1.7	1	3.0	0.38
Attorneys	40	9.2	9	15.3	5	15.2	2.93
College professors	14	3.2	7	11.9	2	6.1	9.57**
Clergymen	8	1.8	1	1.7	0	0.0	0.62
Teachers (high and elementary)	9	2.1	5	8.5	1	3.0	7.63*
Others	8	1.8	3	5.1	1	3.0	2.44
Total	90	20.8	26	44.1	10	30.3	16.27**
Arts and Entertainment	22	5.1	0	0.0	0	0.0	4.88
Grand total	433	100.0	59	100.0	33	100.0	

Note.—Chi squares based on a 2 × 3 classification. The overall association of the whole table yields a $\chi^2 = 44.06$ and $p < .01$.
* $p < .05$.
** $p < .01$.

“primary focus of activity” in the occupation and is considered to be related to most factorizations of interest. The groupings are as follows: Service, Business Contact, Organization, Technology, Outdoor, Science, General Cultural, and Arts and Entertainment. In certain instances, subdivisions within these general groups were established for the more frequent occurring individual occupations. None of the individuals in our series fell into the category of Service.

RESULTS

Smokers versus Nonsmokers. From the data in Table 1, it is apparent that smokers are significantly differentiated from nonsmokers in the frequencies to which they fall into certain occupational classifications. Smokers are significantly more heavily rep-

resented in Business Contact (promoters, salesmen, retail and wholesale dealers, buyers, etc.) and in the field of Organization as a whole, as well as in its subdivision of business executives of all types. Nonsmokers, on the other hand, are significantly more heavily represented in the General Cultural group consisting of attorneys, college professors, clergymen, teachers of high and elementary schools, and especially in the “other” general cultural subdivision (journalists, librarians, museum curators, etc.).

No significant differences are apparent between smokers and nonsmokers in the Technology, Outdoor, Science, and Arts and Entertainment groupings, with the possible excep-

tion of surgeons. Nonsmokers have more than twice the proportion of surgeons than smokers, with this difference closely approaching the 5% level of significance.

Forms of Smoking. The division of smokers into exclusive groupings of pure cigarette, pure pipe, and pure cigar smokers reveals statistically significant variations with regard to Organization and General Cultural occupational groupings (see Table 2). The Organization classification, exemplified by the major and minor executive subdivisions, displays an excess of cigar and cigarette smokers over those expected, and a deficiency

of pipe smokers. The General Cultural group exhibits a particularly large excess of pipe smokers and fewer than expected proportion of cigarette smokers. The deficiency of pure cigarette smokers in the General Cultural classification is to be seen in most of the specific occupational subdivisions, but is especially marked for college professors and teachers because of their strong preferences for the pipe. Apart from the tendency for research scientists to lean toward the pipe and cigar, and the subjects in the field of Arts and Entertainment to smoke cigarettes only, the remaining occupational classifications are

TABLE 3
OCCUPATION GROUPS: ACCORDING TO AMOUNT OF CIGARETTE SMOKING

Group	Cigarettes per day								Total No.	χ^2
	Less than 10 daily		$\frac{1}{2}$ -1 pack daily		1-2 packs daily		2 + packs daily			
	No.	(%)	No.	(%)	No.	(%)	No.	(%)		
Business Contact	10	9.3	18	12.6	34	13.1	6	11.8	68	1.09
Organization										
Major and minor executives	22	20.6	41	28.5	84	32.4	24	47.1	171	12.29**
Bankers	2	1.9	4	2.8	11	4.2	1	2.0	18	1.84
Total	24	22.4	45	31.2	95	36.7	25	49.0	189	12.84**
Technology	2	1.9	11	7.6	25	9.7	1	2.0	39	9.24**
Outdoor	0	0.0	1	0.7	4	1.5	0	0.0	5	2.71
Science										
Research scientists	14	13.1	8	5.6	13	5.0	0	0.0	35	12.67**
Physicians	12	11.2	21	14.6	23	8.9	5	9.8	61	3.18
Surgeons	2	1.9	2	1.4	7	2.7	1	2.0	12	0.81
Total	28	26.2	31	21.5	43	16.6	6	11.8	108	6.81
General Cultural										
Cultural administrators	8	7.5	2	1.4	6	2.3	2	3.9	18	8.56*
Attorneys	13	12.1	12	8.3	27	10.4	4	7.8	56	1.30
College professors	8	7.5	3	2.1	9	3.5	1	2.0	21	5.72
Clergymen	1	0.9	3	2.1	3	1.2	1	2.0	8	0.91
Teachers (high and elementary)	6	5.6	4	2.8	3	1.2	0	0.0	13	8.00*
Others	4	3.7	5	3.5	3	1.2	0	0.0	12	4.83
Total	40	37.4	29	20.1	51	19.7	8	15.7	128	16.20**
Arts and Entertainment	3	2.8	9	6.3	7	2.7	5	9.8	24	7.32
Grand total	107	100.0	144	100.0	259	100.0	51	100.0	561	

Note.—Chi squares based on 2 × 4 classifications. The overall association of the whole table yields a $\chi^2 = 70.77$ and $p < .01$.
* $p < .05$.
** $p < .01$.

not especially distinguished as to the form in which they use their tobacco.

Pure cigarette smoking may be said to be especially characteristic of smokers in Business Contact, Arts and Entertainment, and among physicians and surgeons. This form of smoking is less likely to be found among lawyers, research scientists, college professors, elementary and high school teachers, and others in cultural pursuits. A particular predilection for pure pipe smoking is seen among research scientists, college professors, and elementary and high school teachers. Occupational categories showing a lesser preference for the pipe include physicians and surgeons, executives, and those in arts and entertainment. Cigar smoking, when practiced, is the special preference of the business executive, research scientists, and lawyers.

Rate of Cigarette Smoking. Table 3 gives the occupational distribution of the subjects in accordance with rate of cigarette smoking. It is strikingly apparent that these data indicate a number of consistent, graded sequence of arrangements of the occupational frequencies according to degree of tobacco consumption. Thus, with regard to the Organization group, there is a progressive increase of persons in this classification as the amount of cigarette smoking increases (from 22.4% among those smoking less than 10 cigarettes daily, progressively to 49.0% in smokers of 2 or more packs a day). Conversely, in Science and General Culture groups there is a consistent graded decrease in occupational frequencies with increased amount of smoking. In Business Contact and Technology there is a consistent graded increase of these occupational groupings with increased cigarette consumption, except in the case of the 2 or more packs a day smokers. Graded sequential arrangements are also found for some of the subdivisional occupations of the primary group classifications.

The differentiations here observed may be conveniently expressed as follows. There is a tendency among people in Organization occupations, particularly among major and minor executives, to be heavier rather than lighter cigarette smokers. The opposite tendency, namely, for lighter rather than heavier cigarette smoking, is observed for the Science

and General Cultural groups; and specifically among research scientists, physicians, high and elementary school teachers, and in other general cultural occupations (journalists, librarians, etc.). Persons in Technology are prominent in their predisposition for smoking 1-2 packs a day and their avoidance of the extremes of light or very heavy cigarette smoking. Practitioners of Arts and Entertainment tend to be either very heavy smokers (2+ packs a day) or smokers of $\frac{1}{2}$ -1 pack a day.

DISCUSSION

The relationship between occupation and smoking is a highly complex matter. Involved are such factors as relative ability to purchase tobacco, opportunities for smoking, acceptance of smoking in various kinds of work and situations, social and cultural pressures and patterns, as well as psychological and personality elements operating within the individual. Previous studies dealing with occupation and smoking primarily reflect the influence of social and economic forces. The most extensive survey has been reported by Haenszel, Shimkin, and Miller (1956) in a nationwide investigation of tobacco smoking patterns in the United States. Utilizing the occupational classification of the United States Bureau of the Census, these observers found the highest proportion of nonsmokers in farmers, farm managers, and in professional and technical workers; and the greatest frequency of smokers occurred among craftsmen, foremen, sales workers, and factory operatives. Furthermore, it was reported that there was some evidence in the data of an ordering by social class, as well as strong suggestions of the role of other noneconomic factors in the determination of smoking rates. Similar surveys in England emphasized social and economic distinctions between smokers and nonsmokers, and in the intensity of the smoking habit.

In contrast, the Roe classification used in the present study is designed to mirror psychological and personality influences and to indicate some aspects of self-image. It has been already noted that being a classification by primary focus of activity, it bears relation to factorizations of interest. Signally, Roe

points out that the different occupation groups in the classification show differences in personality pictures "which are in accord with the basis of interests on which they are largely separated."

It is, therefore, of considerable import that the results of the analysis of our data indicate the existence of statistically significant occupational differentiations between smokers and nonsmokers, between the different forms in which smoking was adopted, and in accordance with daily rate of cigarette consumption. Since, if indeed these particular forms of occupational differentiation do in fact reflect some general relationship to factors of personality, then it becomes reasonable to conclude that smoking behavior itself is associated with the individual's personality. The existence of personality differences between occupational groups is being explored in the psychological literature, which is well reviewed by Roe (1956).

To what extent our occupational groupings have genetic implications is difficult to ascertain. According to Dobzhansky (1963):

The occupational groups in a society become genetically meaningful in proportion to the freedom of social mobility which a society provides to its members . . . equality of opportunity appears to lead to formation of genetically specialized occupational classes [p. 113].

It could be argued that our Harvard graduates (by virtue of such status the institution provides and the nature of the student body itself) have had a sufficient measure of equality of opportunity, freedom of choice, and social mobility to satisfy adequately the essential conditions proposed by Dobzhansky. If this be the case, then, the occupational class differentiations here observed indeed may have genetic implications. At any event, this interesting supposition remains a matter for serious speculation.

The above observations are not intended in any way to minimize the role of occupational environment (and other factors) in influencing smoking or nonsmoking practices, the form of smoking adopted, and even the intensity of the smoking habit. It is generally recognized that certain occupations seem to regard one or another form of smoking as

symbols of identification or status: the cigar in business, the pipe-smoking professor or research scientist, the cigarette in journalism, arts, and entertainment. In addition there is the aspect of the smoker's self-image and the influence of the image of him by others, as well as how the individual sees others and how others think he sees himself. Certainly the time factor in terms of opportunity is not irrelevant, as, for example, in the arts and entertainment field, where the pressure of the activity makes it difficult to indulge in the leisure of the preparation, handling, and accompanying rituals of pipe smoking. Nevertheless, the constitutional element is not to be ignored, as may be seen from the penetrating review of anthropological aspects of this subject by Damon and McFarland (1955).

Our findings of occupational variations with smoking habits are to some extent referable to differential personality structures and conceivably of genetic relevance. These findings reinforce the evidence of other studies which indicate that smoking behavior is significantly influenced by host factors within the individual and hence represents manifestations of distinctive constitutional make-ups. The implications of this statement relative to the problem of smoking with health and disease are not inconsequential.

REFERENCES

- DAMON, A., & MCFARLAND, R. A. The physique of bus and truck drivers: With a review of occupational anthropology. *Amer. J. phys. Anthropol.*, 1955, 13, 711-742.
- DOBZHANSKY, T. Genetics and equality. *Science*, 1963, 137, 112-115.
- HAENZEL, W., SHIMKIN, M., & MILLER, M. F. Tobacco smoking patterns in the United States. *U. S. Dept. Hlth. Monogr.*, 1956, 45(Publ. No. 426).
- HEATH, C. W. Differences between smokers and nonsmokers. *Arch. int. Med.*, 1958, 101, 377-388.
- MATARAZZO, J. D., & SASLOW, G. Psychological and related characteristics of smokers and nonsmokers. *Psychol. Bull.*, 1960, 57, 493-513.
- ROE, ANNE. *The psychology of occupations*. New York: Wiley, 1956.
- SELTZER, C. C. Why people smoke. *Atlant. Mon.*, 1962(Jul), 41-44.
- SELTZER, C. C. Morphologic constitution and smoking. *J. Amer. Med. Ass.*, 1963, 183, 639-645.

(Received January 21, 1963)

DETERMINING WORKER PREFERENCES AMONG EMPLOYEE BENEFIT PROGRAMS ¹

STANLEY M. NEALEY

University of California, Berkeley

The paired comparison method was applied to measure the preferences of 1133 members of a trade union among a pay raise, a union shop proposal, a vacation plan, a shorter work week, hospital insurance, and a pension increase. Hospital insurance was most preferred while the shorter work week was least preferred. Differences in preference were markedly related to age and seniority, moderately related to physical-clerical job type, marital status, and number of dependent children. Preference for the pay raise was scarcely related to the demographic variables. The preference judgments were highly transitive and allow the 6 compensation options to be ranked in an ordinal scale.

The employee-attitude survey has often been pressed into service to answer questions about employee preferences among compensation and benefit programs. For instance, the National Industrial Conference Board (1947) conducted a large survey in which workers were asked to rank the most important items from a list of 71 morale factors. "Job security" was ranked first, "compensation" was third, and "vacation and holiday practices" were ranked seventh in importance. The National Retail Dry Goods Association (Fosdick, 1939) asked 3,000 employees in a nation-wide sample to rank eight morale factors. "Credit for all work done" came out first in importance while "pay" was third and "job security" was last.

That information of this kind is nearly impossible to use as a guide in designing employee compensation and benefit programs is a standard complaint.

Kornhauser (1944, p. 132) in a penetrating analysis of the shortcomings of employee-attitude surveys wrote:

The percentage of employees who are pleased or displeased about a particular condition obviously gives no indication of the importance or significance of those feelings as determinants of general satis-

faction. If 75% of a group express dissatisfaction with lockers or dirty windows while only 25% say they are dissatisfied about wage rates, it certainly does not mean that management is justified in

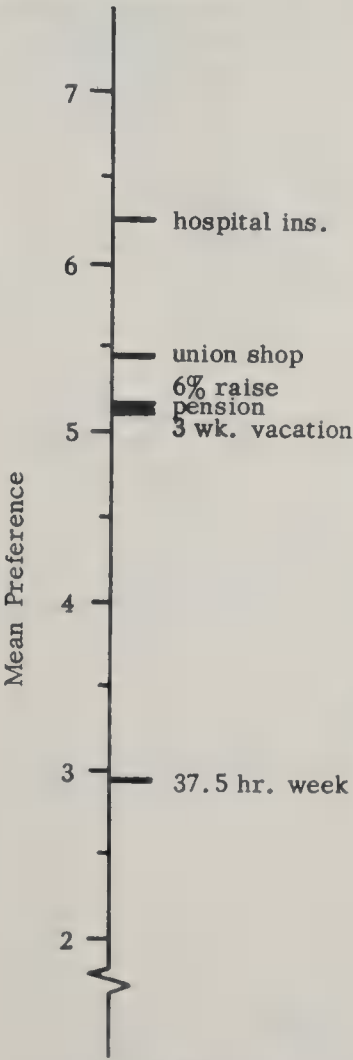


FIG. 1. Relative preference for pay, union shop, and four fringe benefits.

¹ Part of this study was supported by a grant from the General Electric Company and the Institute of Industrial Relations at the University of California, Berkeley. Thanks for financial support and cooperation is also due the staff and members of Local 1245 of the International Brotherhood of Electrical Workers (AFL-CIO). Finally, my thanks to Mason Haire for his continuing help and concern with this research topic.

giving attention to the former and ignoring the latter. The important additional problem is how urgently the expressed attitudes are felt and how influential they are in determining the overall orientation of the individual.

The business magazine *Modern Industry* (1950) took a more direct approach to benefit preference measurements. Setting out to determine the amount of worker interest in pensions, they asked the question: "If you had to choose between a pension at 65 or any one of the following benefits which would you rather have?" and found that pension was preferred over higher pay, sick leave, accident or health insurance, paid vacation, and life insurance in that order. Unfortunately, the form of the question may have

biased these results in favor of the pension, and inferences about preferences among the other five benefits are difficult to make.

Even with these shortcomings the more direct approach taken in the *Modern Industry* survey seems better suited than the employee-attitude survey to the task of determining employee preferences among benefit plans. Surprisingly enough, there seems to be no other published reference to the use of this direct method. The present study uses the paired comparison method to approach preferences directly. It is hoped that it may be a step toward securing valid preference information in a usable form, and a step toward the identification and measurement of the determinants of preference.

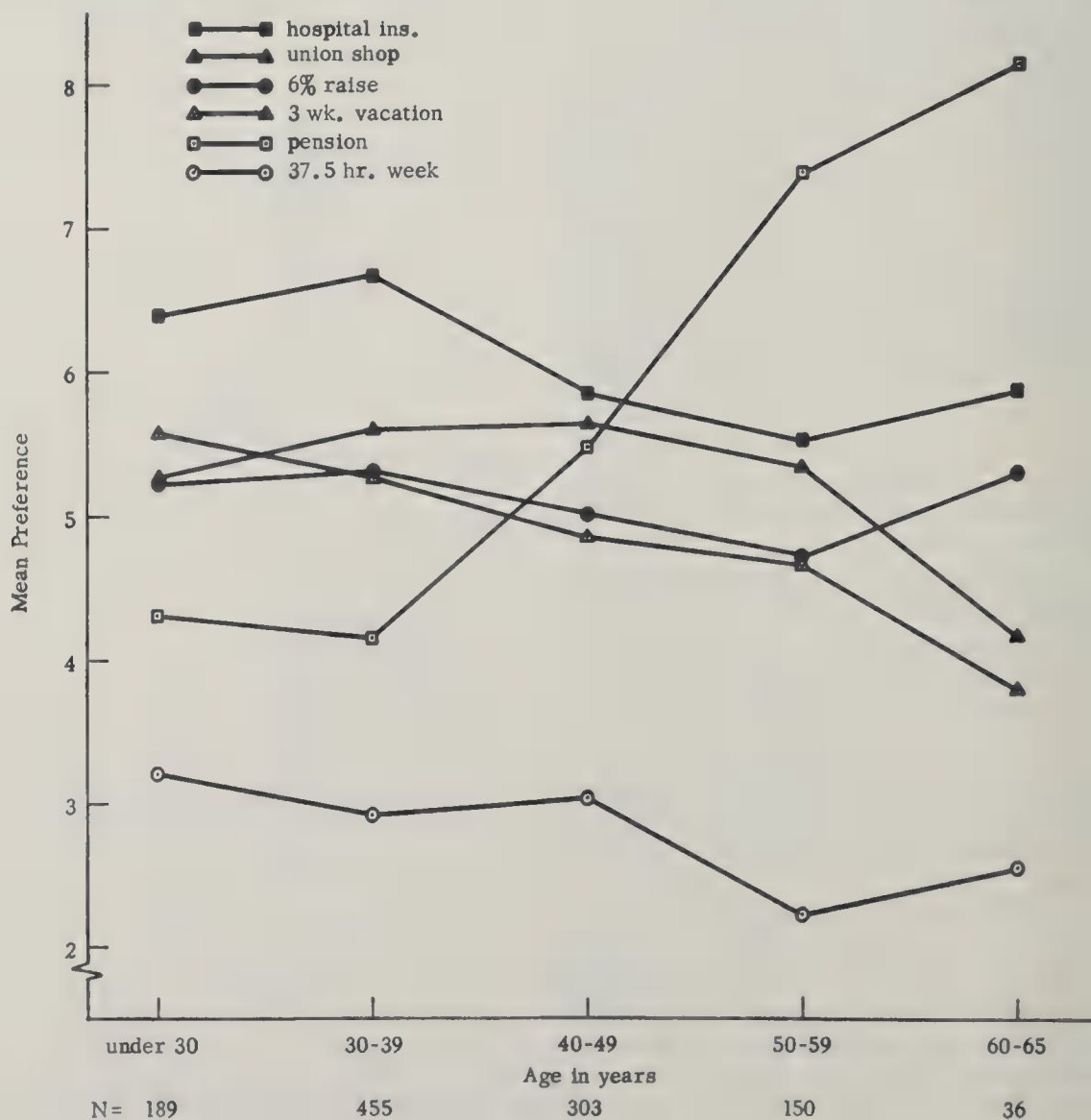


FIG. 2. Benefit preference by age.

METHOD

Questionnaires were sent by mail to nearly 8,300 male members of a Northern California local of the International Brotherhood of Electrical Workers (AFL-CIO). The questionnaires were mailed by the union and returned anonymously to it. Correctly completed questionnaires were obtained from 1,133 males. The questionnaire consisted of six alternative payment and benefit plans presented in a paired comparison format. Each plan was paired with every other plan two times, giving 30 pairs in all. Subjects were asked to mark the plan they preferred out of each pair. The pairs were presented first in one order and then in the other order to control for time-order error. The six plans were listed in their complete form at the head of the questionnaire, but were abbreviated for paired comparison presentation.

The six plans in their complete form are listed below:

1. The Company will pay the cost of an additional \$50 a month pension to be added to my retirement benefits.
2. I will get a 6% raise.
3. The normal work week shall be cut to 37½ hours without any reduction in weekly earnings.
4. The Company will pay the entire cost of full hospital insurance for myself and my family.
5. It will be agreed that all regular employees must be members of the union. (The Union Shop will be put into effect.)
6. I will have 3 weeks' paid vacation a year in addition to my present vacation; the extra vacation to be taken when I choose.

These six plans were chosen so as to be of approximately equal cost to the employer. This aim seems well realized in the cases of the vacation plan, the shorter work week, and the pay raise. The pension plan and the hospital plan may also be close to the other plans in cost, but the actual values involved are somewhat dependent on situational factors. The union shop proposal is of unknown cost.

Six categories of "face" data were included on the questionnaire. These were: age, company seniority, rural or urban place of residence, job title or income, marital status, and number of dependent children. The results to be reported are analyzed in terms of these classifications.

RESULTS AND DISCUSSION

Figure 1 shows the preference dispersion of the total group of 1,133 men. The data are expressed in terms of the mean number of times each plan was chosen by each subject. Since each plan appeared on the questionnaire in 10 pairs, it could be chosen by each subject from 0 to 10 times. Hospital insurance was

clearly most preferred while the union shop was second in preference. Next came the 6% raise, pension, and the 3-week vacation, all virtually equal in preference. The shorter work week was a very poor last in preference. It was chosen less than 30% of the time.

The seven variables we have examined proved to vary widely in the nature and degree of their relations to benefit preferences.

Age appeared to be an active determinant of preference (see Figure 2). Although pension preference rose sharply with age, the effect did not occur until about age 40. Preference for the remaining options, particularly the vacation plan, tended to decrease with age, although the union shop lost support only among the oldest age group. Note also that preference for hospital insurance reached its highest point in the 30-39 year age group, indication that this group felt the greatest need for family medical protection.

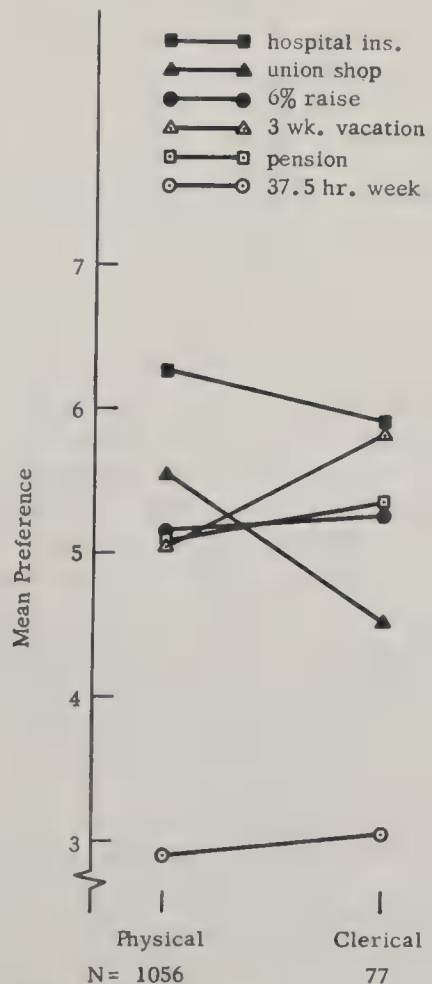


FIG. 3. Benefit preference by physical and clerical job types.

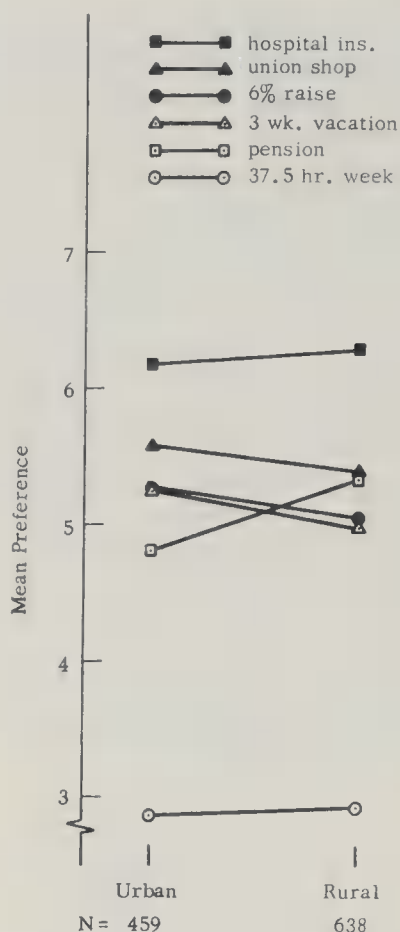


FIG. 4. Benefit preference by urban and rural places of residence.

The curves that resulted from a breakdown of preferences by company seniority were so nearly identical to the age curves that the seniority curves will not be presented.

Job type (clerical workers versus "physical" workers) was a determinant of preference for vacation and the union shop (see Figure 3). Preferences for the other four options were largely unrelated to job type, as the level lines in Figure 3 indicate. Clerical workers showed much higher interest in vacation than did the physical workers. Just the reverse was true with the union shop proposal, a finding that ties in nicely with the traditional lag in unionization of clerical workers.

Rural-urban place of residence, as Figure 4 shows, was of only slight importance as a preference determinant. The pension plan was moderately more attractive to the rural group than the urban group. This difference was independent of age since the mean age of the two groups was nearly identical.

Figure 5 shows the preferences of 1,059 employees in "physical" jobs that had been dichotomized as having high or low skill requirements. Skill level was the only independent variable examined that failed to show any marked changes in preference. This seems quite surprising since skill level is highly correlated with income level, a variable which a priori might be expected to influence preference for several of the plans presented.

As Figure 6 shows, marital status was dramatically related to preference for hospital insurance, pointing up the importance of family responsibility as a preference determinant of this benefit. Less easily predicted was the relatively higher preference among the unmarried group for the shorter work week.

The variable of dependent children, Figure 7, was related to changes in preference for hospital insurance and the pension plan. As might be expected, hospital insurance was

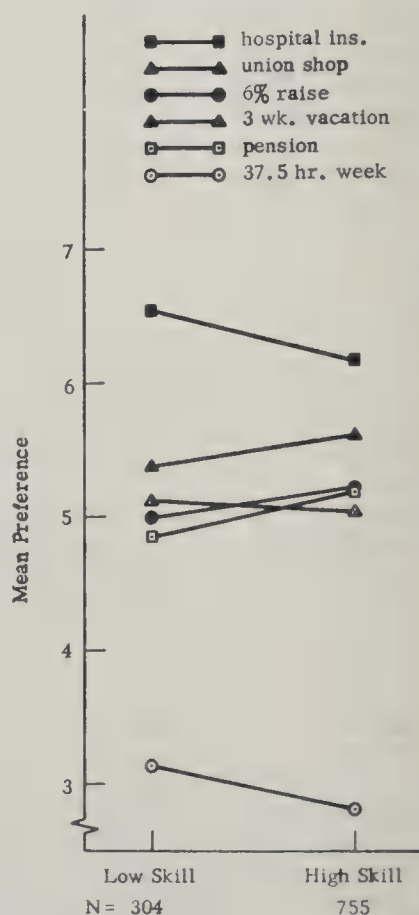


FIG. 5. Benefit preference by high or low skill job classifications.

more attractive to the group having children. Preference for pension among the group without dependent children may be partly related to the fact that this group had a mean age several years higher than the groups with dependent children. There was little difference between the group with one or two children and the group with three or more children.

Shifting our attention from the subject variables to the benefit plans, it may be seen that some, for example, the union shop proposal, hospital insurance, and the pension plan, show striking preference changes with at least two of the subject variables. This is an indication that these plans did not have universal appeal, but rather were more popular or less popular with certain special interest groups.

By contrast, preference for the pay raise, while not outstandingly high relative to other compensation options, remained remarkably constant. Preference for pay did not change

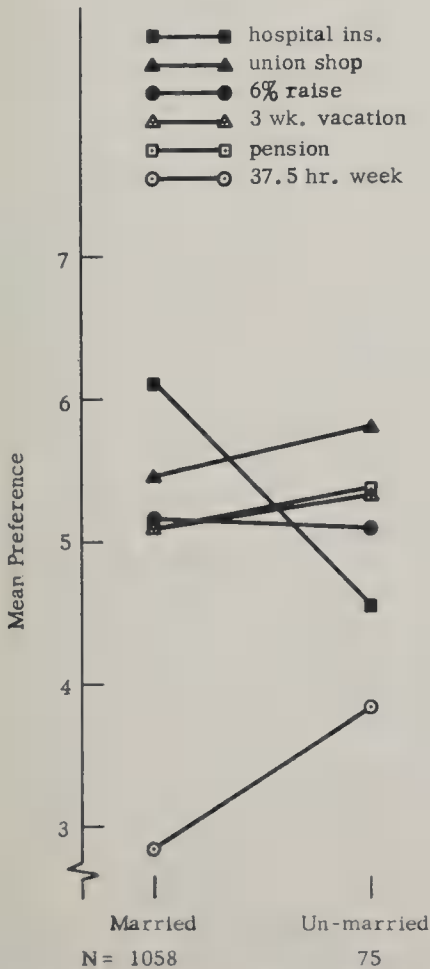


FIG. 6. Benefit preference by marital status.

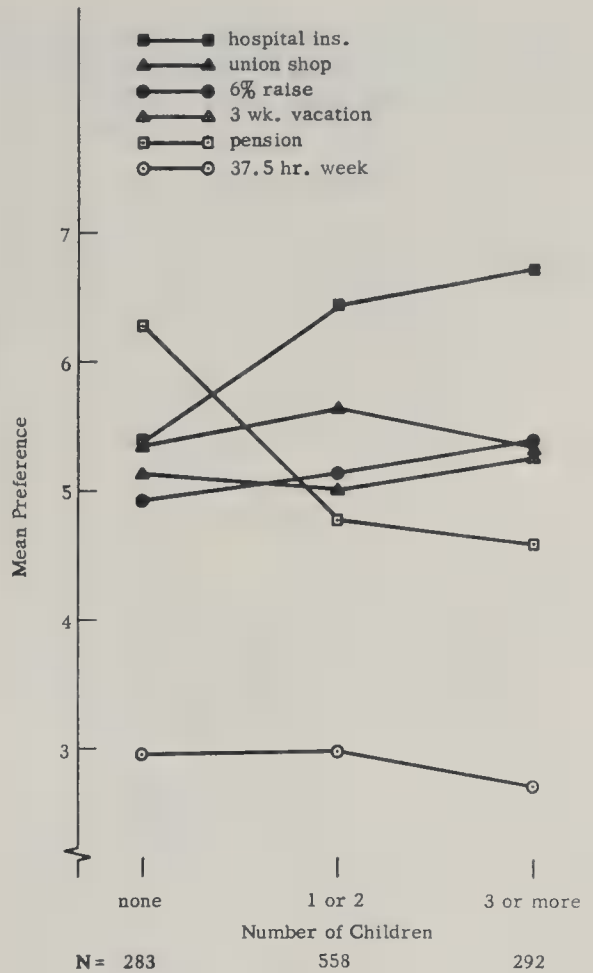


FIG. 7. Benefit preference by number of dependent children.

noticeably with regard to any of the six subject variables.

It should be clear that the determinants of preference for pay and benefits are extremely complex. In interpreting the present findings, care should be taken in generalizing beyond the present situation. While some of the present findings may be common to many industrial situations, others are no doubt unique to this one.

The use of the paired comparison method to obtain preference information about pay and benefits permits the ordinal scaling of preferences, thus improving upon a major difficulty of the earlier approaches.

Torgerson (1958, pp. 27-29) emphasizes transitivity of judgments as the only testable criterion for an ordinal scale. The present paired comparison preference judgments were tested for transitivity within each of the 120 possible triads of benefit plans. Of the 33,990

judgments involved, 96.97% were transitive. It may thus be concluded that the preference attributes of these pay and benefit plans are fundamentally measurable on an ordinal scale.

REFERENCES

- FOSDICK, S. J. Report to the 1939 Convention of the National Retail Dry Goods Association. In G. W. Hartmann & T. Newcomb (Eds.), *Industrial conflict*. New York: Cordon, 1939. Pp. 114-124.
- KORNHAUSER, A. Psychological studies of employee attitudes. *J. consult. Psychol.*, 1944, 8, 127-143.
- MODERN INDUSTRY. Workers sound off on pensions. *Mod. Industr.*, 1950, 19(2), 38-41.
- NATIONAL INDUSTRIAL CONFERENCE BOARD. *Factors affecting employee morale*. (Studies in Personnel Policy No. 85) New York: NICB, 1947.
- TORGERSON, W. S. *Theory and methods of scaling*. New York: Wiley, 1958.

(Received December 17, 1962)

MAINTENANCE OF VIGILANCE IN AN AUDITORY MONITORING TASK¹

J. S. KIDD² AND ANGELO MICOCCHI³

Ohio State University

4 levels of critical signal frequency and 3 levels of task complexity were compared for their effect on vigilance in an auditory monitoring task. Proportionate omission errors increased as the frequency of signals decreased, as expected. However, complexity (defined as the number of categories of critical signals) had an unexpected effect in that relatively poor performance occurred with increased complexity. The results were interpreted as suggesting caution in the use of artificial signals as a means to overcome loss of vigilance in monitoring tasks.

One of the consequences of automation is the generation of a class of tasks which require an operator to monitor a series of signals over extended periods of time during which critical events occur infrequently. Such tasks are common in military operations (i.e., radar monitoring in air defense stations) and are increasing in number in industrial operations (i.e., controlling an automatic petroleum cracking plant). It has been demonstrated repeatedly that continuous monitoring of infrequent-critical-event sequences leads to a loss of vigilance (Mackworth, 1950).

There seems little question that the loss of vigilance process is based on the decay of a perceptual set. Presumably, the preparatory responses which comprise a perceptual set must be reinforced to be maintained. This appears to be an instance of the venerable law of disuse. However, such broad explanations have little utility for the practical matter of developing means to sustain alertness under adverse task conditions. Reinforcement theory provides more detailed concepts. Some questions which are suggested by rein-

forcement theory have direct relevance. For example, the following question may be posed: Can alertness be sustained by increasing the complexity of the signal discrimination aspect of the task (thereby giving the covert elements of the task more variety) or is the level of alertness primarily dependent on the actual frequency of critical signals (which, by definition, elicit an *overt* response)? Garvey, Taylor, and Newlin (1959) have found that the use of artificial signals has a beneficial effect on vigilance in a visual task setting. This is an instance of increased complexity in the sense that the subject is required to detect all signals and to discriminate between real and artificial signals.

The present experiment was designed to shed some light on this issue. It was also intended to broaden the content of vigilance research by utilizing the auditory mode of presentation in the context of a "classical" vigilance task (Frankmann & Adams, 1960). The hypothesis under evaluation was that an increase in monitoring task complexity, independent of critical signal frequency, can result in an improvement in vigilance or reduce the loss-of-vigilance effect.

METHOD

The task imposed on the subject was to monitor a continuous stream of auditory messages, a few of which required positive identification and an overt recording response. These latter messages were designated as "critical" in the series.

The content of the message sequence was a series of synthetic stock market quotations. Sixty company names were used in a 2-hour sequence consisting

¹ This research was carried out in the Laboratory of Aviation Psychology and was supported by the Air Force Systems Command, United States Air Force under Contract No. AF 33(616)-6166, monitored by the 6570th Medical Research Laboratories. Permission is granted for reproduction, translation, publication, use, and disposal in whole and in part by or for the United States Government.

The authors appreciate the help and guidance provided by George E. Briggs and George Wright.
² Now Principal Staff Scientist, Aircraft Armaments, Incorporated, Cockeysville, Maryland.

³ Now with Dunlap and Associates, Palo Alto, California.

of 2,400 quotations or 1 every 3 seconds. The order of firm names in the list was random. The subject was assigned a company name (or names depending on the experimental condition), and was instructed to record on a standard answer sheet the price quoted for this company each time it was mentioned.

During test trials, the subject was seated in a small, well-lighted and well-ventilated room. The stock price sequence was played back from a standard tape recorded through an Air Force muff-style head set.

One hundred twenty male subjects were recruited from the undergraduate population at large. They were paid \$1.00 an hour for their time. An incentive bonus of \$.50 was paid to the subjects if they completed the task and made no errors. Answer sheets were scored immediately after each test trial so prompt payment of both the fixed wage and the incentive bonus could be made.

Two task factors were varied: the total frequency of critical messages and the number of different firm names designated as critical. The frequency factor was sampled at four levels: an average rate of (a) one critical message every 60 seconds, (b) one critical message every 180 seconds, (c) one critical message every 360 seconds, and (d) one critical message every 720 seconds. (This variation was implemented by having different firms appear on the tape program with different frequency and assigning firm names to subjects corresponding to the appropriate condition designation.) The task complexity variable (the number of different firm names assigned) was sampled at three levels: one, two, or three firms to be identified and their prices reported.

The two variables were organized factorially to give 12 unique conditions. Ten subjects were observed under each condition.

Since frequency of occurrence was an element of both variables, explicit confounding was avoided by equal allocation of critical signals to identification classes. For example, under the two classes, one critical message per 60 seconds condition; one class would generate one message every 120 seconds on the average and the other class would also generate one message every 120 seconds. Thus, the overall average would remain at one critical message every 60 seconds.

RESULTS

Commission errors (false positive responses) were relatively rare. Consequently, the relationship between commission error frequency and critical message frequency is somewhat erratic. However, the general trend is toward a decrease in commission errors as critical message rate is decreased. If subjects are classified into those who did, as opposed to those who did not, make errors of commission and this dichotomy is organized across

TABLE 1
DISTRIBUTION OF SUBJECTS UNDER FOUR CONDITIONS OF CRITICAL MESSAGE FREQUENCY IN TERMS OF ERRORS OF COMMISSION

Number of commission errors per trial	Average interval between critical messages			
	60 sec.	180 sec.	360 sec.	720 sec.
0	14	20	16	28
1+	16	10	14	2

the frequency of critical message factor, a chi square analysis yields a value of 16.84 which is significant at the .001 level of confidence. The data is presented in matrix form in Table 1. The inconsistency of the trend, however, makes any interpretation somewhat suspect.

The treatment of omission errors must be modified by a consideration of the relative opportunities to make such errors. Therefore, the error frequencies were transformed into a percentage score before statistical processing. The average proportion of omission errors as a function of frequency of critical messages is presented in Figure 1. The regular drop in error rate as critical message frequency increases is in the expected direction.

This same effect is verified by a different mode of data treatment. The distribution of omission error scores had a natural break near the middle of the range (2 errors per hundred signals). If the subjects are dichotomized by condition on either side of the natural break, a contingency table is obtained

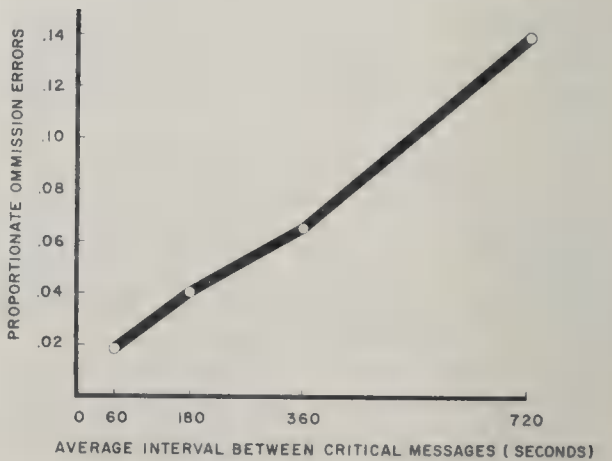


FIG. 1. Proportionate omission errors as a function of the average interval between critical messages.

TABLE 2

DISTRIBUTION OF SUBJECTS UNDER FOUR CONDITIONS OF CRITICAL MESSAGE FREQUENCY USING A PROPORTIONAL OMISSION ERROR SCALE

Proportionate omission errors	Average interval between critical messages			
	60 sec.	180 sec.	360 sec.	720 sec.
0-.02	22	20	15	11
.03-.83	8	10	15	19

which is presented in Table 2. Chi square analysis of the frequencies gives a value of 10.05 which is significant beyond the .02 level of confidence ($df = 3$). From the Table, it can be seen that this contingency relationship is free of reversals.

The most pronounced outcome was the increase in omission errors with the increase in the number of identification categories. When one firm name was designated, the error rate was less than one per subject per trial. When three firm names were critical, an average of almost three errors per subject per trial was obtained.

TABLE 3

DISTRIBUTION OF SUBJECTS UNDER THREE CONDITIONS OF DETECTION CATEGORY COMPLEXITY

Omission error frequency	Number of critical message categories		
	One	Two	Three
0	20	13	13
1-2	17	20	10
>2	3	7	17

When score distributions were prepared, it was apparent that the data could be readily trichotomized without violating the assumptions of the chi square test. Again, natural grouping of the scores was used as a guide. A contingency table was prepared which is reproduced as Table 3. Chi square analysis resulted in a value of 17.15 which is significant at the .01 level of confidence ($df = 2$).

DISCUSSION

The results obtained tend to add weight to the contention that the occurrence of an overt response plays a crucial role in the maintenance of attention in a monitoring task. The original hypothesis that discriminative complexity would produce a comparable effect was refuted. The data show just the opposite outcome. The more varied the task, in the sense that multiple detection requirements constitute variety, the more frequent were errors. This conclusion would weigh against the modification of monitoring tasks by the insertion of false, spurious, or irrelevant signals.

REFERENCES

- FRANKMANN, J. P., & ADAMS, J. A. Theories of vigilance. *USAF CCDD tech. Note*, 1960(Apr.), No. 60-25.
- GARVEY, W. D., TAYLOR, F. V., & NEWLIN, E. P. The use of "artificial signals" to enhance monitoring performance. *USN Res. Lab. Rep.*, 1959, No. 5629.
- MACKWORTH, N. H. Researches on the measurement of human performance. *Med. Res. Council spec. Rep. Ser.*, 1950, No. 268.

(Received December 27, 1962)

PREDICTION OF ACADEMIC ACHIEVEMENT USING THE EDWARDS NEED ACHIEVEMENT SCALE

JERALD G. BACHMAN¹

University of Pennsylvania

A study of the utility of the Edwards need Achievement scale (*n Ach*) as (a) a supplement to academic aptitude tests, and (b) a predictor of over- and underachievement. Ss were 61 male college sophomores. A correlational analysis was carried out among the following measures: Edwards *n Ach*, Scholastic Aptitude Test (SAT), grade-point average (GPA), and a derived measure of over- and underachievement. The results indicated (a) no increment in prediction of GPA when *n Ach* scores were added to SAT scores in a multiple regression equation, and (b) little success in predicting over- and underachievement from *n Ach* scores. Implications for the construct validation of the Edwards *n Ach* scale were discussed. The use of individual course grades as alternative criteria of academic achievement was explored.

Following the appearance of the Edwards Personal Preference Schedule (Edwards, 1954), a number of studies have been undertaken to discover the relationship between academic achievement and certain need scales measured by the Edwards. These studies have provided data relevant to two important problem areas: the prediction of academic achievement, and the construct validity of the Edwards. Since the most consistent positive results have been obtained for the need Achievement (*n Ach*) scale, the present study is limited to an examination of that measure.

In the prediction of academic achievement, the practical utility of a personality measure depends upon its ability to account for a portion of the criterion variance *not predicted by an academic aptitude test*. Weiss, Wertheimer, and Groesbeck (1959), using a sample of 49 undergraduate psychology students, found that when *n Ach* scores were added to academic aptitude test scores, the correlation with grade point average (GPA) was increased from .55 to .64. Using a related procedure, Goodstein and Heilbrun (1962) examined the correlation between *n Ach* and GPA, with academic aptitude scores partialled out. They found correlations of .24 for a

sample of 206 males and .07 for 151 females (undergraduate college students). A somewhat different approach was taken by Gebhart and Hoyt (1958), who found that *n Ach* scores were higher for overachievers than they were for underachievers. (Overachievers were defined as those students whose grades were substantially higher than predicted by academic aptitude test; the opposite condition defined underachievers.) A replication by Krug (1959) yielded very similar results.

The results cited above are relevant also to the construct validity of the Edwards *n Ach* scale. Presumably, achievement motivation will have a positive influence upon academic accomplishment, somewhat apart from the effects of intellectual ability or measured academic aptitude. Accordingly, a valid measure of *n Ach* would be expected to correlate positively with amount of overachievement (and thus negatively with amount of underachievement). Stated differently, *n Ach* should correlate positively with academic achievement, *holding academic aptitude constant*.

The primary purpose of the present study is to provide further evidence concerning the ability of the Edwards *n Ach* measure to supplement standard tests of academic aptitude in predicting academic achievement and to discriminate between over- and underachievers. A secondary purpose is to explore an alternative to GPA as the criterion of academic achievement.

¹ Now at Survey Research Center, University of Michigan.

The author wishes to thank M. S. Viteles, under whose guidance the study was conducted, for his continued interest and encouragement. Thanks are due also to A. Wolfe and R. Mackowiak for providing much of the data for this analysis.

Criterion Problem. One of the perennial problems involved in the study of academic achievement has been the development of adequate criteria. The most frequently used criterion has been the student's GPA, taken over one or more semesters. This criterion is open to criticism, in part because of the inequalities arising when students take varying combinations of courses under different teachers.

If courses differ widely in their level of difficulty, and if instructors assign grades according to different standards, then unwanted variance is incorporated in the criterion. Indeed, a systematic attenuation of the validity coefficient for a predictor will occur if students obtaining high predictor scores characteristically choose the most difficult courses, those in which grading standards are relatively stringent.

This particular type of distortion may be avoided by using as a criterion those grades assigned by one teacher in a single course. To be of greatest practical value, such a procedure would have to be repeated many times with a variety of courses. The secondary purpose of the present study is simply to demonstrate the feasibility of such an approach, in this case using introductory psychology grades as an alternate criterion of academic achievement.

METHOD

Subjects. All subjects were selected from two introductory psychology sections at the University of Pennsylvania. All sophomore males who had participated in a special program of testing as incoming freshmen were included. The subjects in Section A ($N=37$) formed the validation group; those in Section B ($N=24$) served as a cross-validation group.

Measures. The following measures were used in the study:

1. The *n* Ach scale of the Edwards Personal Preference Schedule. The complete Edwards had been administered as a part of a program of testing for all incoming freshmen.
2. Scholastic Aptitude Test (SAT) Total Score. Each student had taken this test prior to application for admission to the University.
3. Freshman year GPA.
4. Total examination points in introductory psychology (Psych). These points were the basis for final course grades.

Measures of Overachievement. In order to gain a more accurate estimate of the ability of *n* Ach

scores to predict overachievement and underachievement, it seemed appropriate to develop a continuum of overachievement (rather than the sort of dichotomy used by Gebhart & Hoyt, 1958, and Krug, 1959). This continuum of overachievement (O-A) was derived by subtracting predicted grades from obtained grades: $O-A = \text{obtained grade} - \text{predicted grade}$. Predicted grades were based upon SAT scores regressed according to the validity coefficients found in the present study. Standard scores were used throughout to simplify computation. Using this formula, if a student's attained grade was higher than predicted, then his O-A score was positive (indicating overachievement). A student falling short of his predicted grade received a negative O-A score (indicating underachievement).²

Since two separate criteria of performance (GPA and Psych) were used in this study, two O-A scores were computed for each subject:

$$\begin{aligned} O-A_1 &= Z_{GPA} - Z_{\text{regressed SAT}} \\ O-A_2 &= Z_{\text{Psych}} - Z_{\text{regressed SAT}} \end{aligned}$$

(It should be noted that since Section B was used as a cross-validation sample, the validity coefficients for Section A served as the regression coefficients in computing all O-A scores.)

Statistical Analysis. Product-moment coefficients of correlation were computed among the measures listed above. In addition, SAT and *n* Ach scores were combined in multiple product-moment correlations with each of the two criteria of academic performance.

RESULTS

Prediction of Academic Performance. The results indicating the contribution of the Edwards *n* Ach scale in the prediction of academic performance are summarized in Table 1. In all cases the correlations between *n* Ach and the criteria of academic performance were positive. Nevertheless, the correlation with GPA did not reach the .05 level of

² The O-A score was originally suggested by the "D score" used by Ghiselli (1956, 1960) in the construction and validation of "predictors of predictability." Ghiselli (1960) defined the D score as "the difference between standard predictor scores, z_p , and the standard criterion scores, z_o [p. 3]." The O-A score differs from Ghiselli's formulation at two essential points: (a) the O-A score retains the sign of the difference (positive or negative) while Ghiselli's D score does not; (b) the standard predictor score is regressed (multiplied by the validity coefficient) in determining the O-A score, but it is not regressed in determining the D score.

It should be noted that it might be possible to develop predictors of over- and underachievement empirically using Ghiselli's approach, substituting O-A scores for D scores.

TABLE 1
CORRELATIONS AMONG PREDICTORS AND CRITERIA OF ACADEMIC PERFORMANCE

	Section A ^a validation sample			Section B ^b cross-validation sample		
	GPA	Psych	SAT	GPA	Psych	SAT
n Ach	.28	.41*	.30	.19	.45*	.50*
SAT	.59**	.77**		.51*	.55*	
SAT + n Ach	.57**	.78**		.47*	.59**	

Note.—In all cases two-tailed tests of significance were used.
^a N = 37.
^b N = 24.
* $p \leq .05$.
** $p \leq .01$.

significance (two-tailed) in each group. On the other hand, in both sections the correlation between n Ach and Psych reached the .05 significance level.

The addition of the n Ach score to the SAT score in a multiple regression equation failed to yield any appreciable change in accuracy of prediction of either criterion.

Differentiation of Over- and Underachievers. Table 2 presents the correlations between n Ach scores and the two measures of overachievement. None of the correlations reached the .05 level of significance.

Criteria of Academic Performance. Tables 1 and 2 indicate that in every instance the use of psychology grades as criteria resulted in higher correlations than those obtained using GPA. However, the two psychology sections differed substantially in this respect.

Section A psychology grades appeared to be much more “predictable” (from SAT scores) than did those of Section B.

DISCUSSION

The results of the present study offer little support for the use of the Edwards n Ach scale as a supplementary predictor of academic achievement. This finding is somewhat inconsistent with that reported by Weiss, Wertheimer, and Groesbeck (1959); however, it should be noted that none of the other studies cited above presented clear evidence that the n Ach scale can substantially improve the prediction of academic achievement.

A consideration of the data dealing directly with the prediction of overachievement suggests that the n Ach scale is of little value in differentiating over- and underachievers. Studies by Gebhart and Hoyt (1958) and Krug (1959) found overachievers to have significantly higher n Ach scores than underachievers; however, these investigators used only extreme groups (about 30% of the population in each case), and even then the distributions of under- and overachievers differed by only about half of one standard deviation. This suggests that the underlying correlations were rather small and of very limited predictive utility.

A note concerning the construct validity of the Edwards n Ach scale may be in order. Some doubt as to the construct validity of this scale derives from its failure to correlate consistently with another scale designed to measure the same variable. Weiss,

TABLE 2

CORRELATIONS BETWEEN EDWARDS NEED ACHIEVEMENT SCORES AND MEASURES OF OVER- AND UNDERACHIEVEMENT

	Section A ^a validation sample		Section B ^b cross-validation sample	
	0-A ₁ score z _{GPA} -z _{SAT} ^c	0-A ₂ score z _{Psych} -z _{SAT} ^d	0-A ₁ score z _{GPA} -z _{SAT} ^c	0-A ₂ score z _{Psych} -z _{SAT} ^d
n Ach	.11	.29	-.14 ^e	.00 ^e

^a N = 37.
^b N = 24.
^c Regressed according to the validation coefficient for Section A (.59).
^d Regressed according to the validation coefficient for Section A (.77).
^e If the validity coefficients for Section B had been used, the indicated correlations would have been -.12 and .14 instead of -.14 and .00, respectively.

Wertheimer, and Groesbeck (1959) found a correlation of only .26 between the Edwards n Ach scale and McClelland's (1953) picture-story measure of achievement motivation. Melikian (1958) and Himmelstein, Eschenbach, and Carp (1958) reported very low and non-significant correlations between the two measures. Cronbach and Meehl (1955) have pointed out the special importance of negative findings and the need for considering them fully in any attempt to demonstrate construct validity. The present lack of support for the n Ach scale as a predictor of academic achievement raises further doubt as to its status as a measure of need for achievement.

A secondary purpose of the present study was to compare two criteria of academic performance. The results suggest the need for more careful consideration of the source of criteria in assessing the predictive potential of tests. The findings concerning criteria are limited, however. There was a substantial difference between the two psychology sections in terms of the "predictability" of their grades. Moreover, there is no way of knowing whether the tendency found in these psychology sections would also hold for classes in English or mathematics, for example. Much further study is necessary before any conclusion can be reached. The present study may indicate the usefulness of such exploration.

REFERENCES

- CRONBACH, L. J., & MEEHL, P. E. Construct validity in psychological tests. *Psychol. Bull.*, 1955, **52**, 281-302.
- EDWARDS, A. L. *Edwards Personal Preference Schedule*. New York: Psychological Corporation, 1954.
- GEBHART, G. G., & HOYT, D. T. Personality needs of under- and overachieving freshmen. *J. appl. Psychol.*, 1958, **42**, 125-128.
- GHISELLI, E. E. Differentiation of individuals in terms of their predictability. *J. appl. Psychol.*, 1956, **40**, 374-377.
- GHISELLI, E. E. The prediction of predictability. *Educ. psychol. Measmt.*, 1960, **20**, 3-8.
- GOODSTEIN, L. D., & HEILBRUN, A. B., JR. Prediction of college achievement from the Edwards Personal Preference Schedule at three levels of intellectual ability. *J. appl. Psychol.*, 1962, **46**, 317-320.
- HIMMELSTEIN, P., ESHENBACH, A. E., & CARP, A. Interrelationships among three measures of need achievement. *J. consult. Psychol.*, 1958, **22**, 451-452.
- KRUG, R. E. Over- and underachievement and the Edwards Personal Preference Schedule. *J. appl. Psychol.*, 1959, **43**, 133-136.
- MCCLELLAND, D. C., ATKINSON, J. W., CLARK, H. A., & DOWELL, E. L. *The achievement motive*. New York: Appleton-Century-Crofts, 1953.
- MELIKIEN, L. H. The relationship between Edwards' and McClelland's measure of achievement motivation. *J. consult. Psychol.*, 1958, **22**, 296-298.
- WEISS, P., WERTHEIMER, M., & GROESBECK, B. Achievement motivation, academic aptitude, and college grades. *Educ. psychol. Measmt.*, 1959, **19**, 663-666.

(Received January 2, 1963)

MAGNIFICATION AS A VARIABLE IN SUBMINIATURE WORK

J. RICHARD SIMON¹

University of Iowa

2 experiments are reported concerning the effects of magnification on the motor skills involved in subminiature work. In 1 study dealing with the interaction of magnification and object size, Ss used a tweezers to pick up spherical metal dots of 2 different sizes under 2 magnifications. In another study dealing with the effects of varying magnification on duration of travel movements under differing precision conditions, Ss performed a repetitive wire positioning operation. In both experiments, Ss viewed their task through a stereoscopic microscope. Results indicated that: the optimum magnification for pickup varied with the size of the object manipulated, there was no interaction between magnification and task precision, and travel movements were slower when performed under higher magnifications.

Perhaps no industrial operation makes more demands on the perceptual and motor capabilities of the worker than that of assembling tiny electronic components. A typical task might involve picking up a part literally the size of a speck of dust or positioning a wire one fifth the diameter of a human hair into a target hole a few thousandths of an inch in diameter. As a result of this trend toward increasing miniaturization, the stereoscopic microscope has become a familiar and necessary fixture on the modern assembly line.

The present study is concerned with the effects of magnification as a perceptual variable on the motor skills involved in subminiature work. What magnification should be employed in performing a given subminiature task? Is it possible to predict on the basis of the size or precision requirements of a task what the appropriate magnification may be? The answers to these questions could be of real importance to our advancing technology.

In the first systematic study of the effects of magnification on subminiature skills, Nayyar and Simon (1963) used a task which consisted of picking up, transporting, and assembling metal dots .010 inches in diameter. Subjects (Ss) performed the task under three magnifications: 20 \times , 30 \times , and 40 \times . One result of their study was that the dura-

tion of the pickup element was significantly shorter when Ss performed the task using 30 \times . The question which arises is: "Is 30 \times the optimum magnification for picking up a range of dot sizes or does the optimum magnification vary with the size of the object to be manipulated?" The first experiment described in this paper was designed to answer this question. Another result of the Nayyar and Simon study was that travel movements were slower when Ss performed the task using the highest power (40 \times). The second experiment described here was designed to study further the effect of magnification on the duration of travel movements within the magnified field and also to investigate the possible interaction of magnification and task precision.

EXPERIMENT I

This experiment was concerned with pickup time as a function of magnification and object size. The Ss' task was to look through a microscope and pick up, with a tweezers, spherical metal dots of two different sizes: .010 inches and .015 inches in diameter. The task was performed under two magnifications: 20 \times and 30 \times .

Apparatus

An American Optical Company stereoscopic microscope was used. A 10 \times eyepiece was combined with either a 2 \times or 3 \times objective to obtain magnifications of 20 \times or 30 \times . Under the field of the microscope was a shallow cylindrical aluminum container $\frac{3}{8}$ inch in diameter and $\frac{1}{4}$ inch high. The container was constructed in two halves which were separated by a nonconducting divider. The right half of the container was used as a receptacle for about 75 metal dots. The left half contained a

¹ The author gratefully acknowledges the assistance of Ravi Nayyar and James Wolf who collected the data, David Hersher who helped with the statistical analysis, and Dee W. Norton who advised on the analysis in Experiment II.

$\frac{1}{16}$ -inch diameter hole into which the dots were placed after pickup. Two Hunter KlocKounters were connected to the right half of the container. One KlocKounter recorded pickup time (in hundredths of a second); that is, it recorded the time the tweezers was in contact with either the right half of the container or with a dot before it was lifted from the container. The second KlocKounter recorded the number of times during the pickup element that the tweezers contacted the dots or the container. When *S* touched a dot or the container surface during the pickup process, a 16.5-volt dc current passed through the tweezers and energized a relay which, in turn, activated the two KlocKounters. A Universal microscope illuminator was adjusted to provide a light intensity of 450 foot-candles to the work area.

Experimental Design and Procedure

Twenty-four right-handed *Ss*, 8 males and 16 females, participated in the experiment. Each *S* performed under four experimental conditions; that is, *S* picked up both sizes of dots (.010 inch and .015 inch) under both magnifications (20 \times and 30 \times). The order of presentation of the conditions was completely counterbalanced, each *S* performing in a different one of the 24 possible orders.

The *Ss* were given 10 familiarization trials during which they became accustomed to moving the tweezers in the microscope field. A familiarization trial consisted of first touching the right half of the container with the tweezers and then the left half and repeating this cycle five times.

On the test trials, a supply of dots, either .010 inch or .015 inch depending on the experimental condition, was placed in the container. The *Ss* were told to "pick up a dot with the tweezers from the right half of the container and place it in the hole in the left half. Repeat this five times as quickly as possible." A trial consisted of transferring five dots into the hole. Six trials were given under each experimental condition, the first three of which were practice trials. Total pickup time and number of contacts made on the last three trials for each condition constituted the experimental data. From these data, the median time and the median number of contacts were computed for each *S* under each experimental condition. While performance during the test trials was reasonably stable, some learning was still taking place as evidenced by a significant ($p < .01$) order effect for both pickup time and number of contacts.

Results

Analyses of variance revealed a significant interaction between magnification and object size in the case of both pickup time ($F = 37.3$; $p < .001$) and number of contacts ($F = 32.1$; $p < .001$). Figure 1 clearly shows that the optimum magnification depended

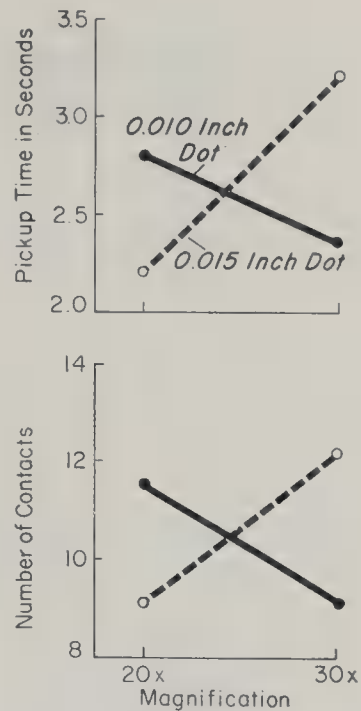


FIG. 1. Pickup time and number of contacts as a function of magnification and object size.

upon the size of the object manipulated; 20 \times was superior for picking up the .015-inch dot while 30 \times was superior for picking up the .010-inch dot. The similarity between the two graphs in Figure 1 suggests the nature of the effect of magnification on the pickup element. It will be noted that the duration of pickup under a given magnification-object size condition was directly related to the number of contacts made.²

Table 1 summarizes the effects of magnification and object size on pickup time and number of contacts per trial. There was an overall significant effect of magnification on pickup time ($F = 5.72$; $p < .05$), with performance under 30 \times being 12% slower than under 20 \times . The effect of magnification on number of contacts was not significant. The

² Data presented for the contacts measure represent the mean of the median number of contacts under each experimental condition. Since, in some instances, the median trial in terms of contacts did not correspond to the median trial in terms of time, the contacts measure was reanalyzed using contacts made during the median trial chosen on the basis of the time measure. Results using this procedure were the same as those reported; that is, they indicated a direct relationship between pickup time and number of contacts.

TABLE 1
EFFECT OF MAGNIFICATION AND OBJECT SIZE ON PICKUP TIME
AND NUMBER OF CONTACTS

	Magnifi- cation	Object size		Mean
		.010 inch	.015 inch	
Pickup time (seconds) per trial	20×	2.80	2.20	2.50
	30×	2.37	3.21	2.79
Mean		2.59	2.71	
Number of contacts per trial	20×	11.50	9.08	10.29
	30×	9.08	12.21	10.65
Mean		10.29	10.65	

main effect of object size was not significant for either pickup time or number of contacts.

EXPERIMENT II

The purpose of this experiment was to investigate the duration of travel movements as a function of magnification and task precision. The Ss' task was to look through a microscope and perform a repetitive wire positioning operation under three different magnifications: 20×, 30×, and 40×. The precision requirements of the task were varied by changing the diameters of the holes into which the wire was positioned.

Apparatus

Subjects viewed the task through an American Optical Company stereoscopic microscope fitted with a 10× eyepiece. Magnifications of 20×, 30×, and 40× were obtained by using 2×, 3×, and 4× objective lenses. Under the microscope was a .004-inch-thick brass sheet which contained two pairs of holes; the two holes used for the high precision task were .021 inch in diameter, while the two holes used for the low precision task were .051 inch in diameter. Each pair of holes was separated by $\frac{1}{8}$ inch from center to center. The holes were oriented perpendicular to S so that the repetitive wire positioning task involved movements toward and away from S rather than movements to the left and right. The brass sheet could be positioned so that either the high precision task or the low precision task was in the magnified field.

Subjects used a Dumont 3C tweezers which had a small piece of .010-inch wire soldered to the tip. The wire was bent in such a way that when the tweezers were held properly, a $\frac{1}{8}$ -inch portion of the wire was nearly vertical. The Ss' task involved positioning the wire into one and then the other of the pair of holes in the brass sheet. Directly beneath each hole was a plastic receptacle containing a pool of mercury. The mercury pools were connected to a motion analyzer (Nayyar & Simon, 1963) so that when the wire was placed in a hole,

it contacted the mercury and activated the timing circuit.

Two Hunter KlocKounters were used to record in hundredths of a second the durations of the travel movements away from and toward S. The KlocKounter which recorded "travel away" time started when S lifted the wire from the near hole and stopped when S placed the wire in the far hole. The KlocKounter which recorded "travel toward" time started when S lifted the wire from the far hole and stopped when S placed the wire in the near hole. At the end of a trial which consisted of five travel movements in each direction, the total times for the two travel elements were read from the two KlocKounters.

Experimental Design and Procedure

Twenty-four right-handed female undergraduates served as Ss. Twelve Ss were randomly assigned to each of the two precision conditions. Each S performed the task under all three magnifications. Four of the 12 Ss in each precision condition were randomly assigned to each of the three orders of the basic 20×, 30×, 40× sequence, thus counterbalancing the effect of order of presentation of the magnification conditions. The design is basically a Type VII (Lindquist, 1953) extended to include a between-subjects factor.

Subjects served in two experimental sessions. The sessions lasted about 40 minutes each and were separated by at least 24 hours. The first session was devoted to familiarization and practice. During this session, Ss performed 10 trials on each magnification in the particular order and precision condition to which they were assigned. A trial consisted of alternately placing the wire in one hole and then the other, repeating this cycle five times. A trial started and ended with the wire in the hole nearest S. Each trial, then, consisted of five travel movements away from S and five travel movements toward S. During the second or test session, Ss performed 15 trials under each magnification condition. The mean of the last 10 trials under each condition was used as the criterion measure. Each S was instructed to

"work as quickly as you can." Analysis of the learning curve indicated that by the start of the test session, Ss had reached a fairly stable level of performance. However, a test of the order effect (see Table 2) revealed that some learning also took place during the test session.

Results

Table 2 summarizes the analysis of variance. The duration of travel was significantly affected by the precision of the task ($F = 22.28$; $p < .001$); that is, travel movements between the small target holes were significantly slower than movements between the large target holes. The main effect of magnification was also significant ($F = 4.30$; $p < .05$). The interaction of magnification and precision was, however, not significant.

Figure 2 shows the duration of travel under

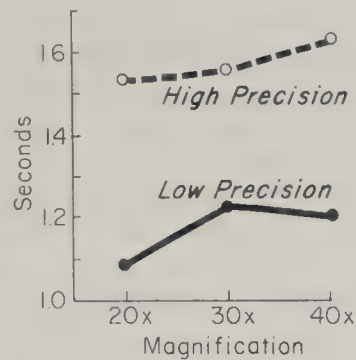


FIG. 2. Effect of magnification on duration of travel for high and low precision tasks.

the three magnification conditions for both the high and low precision tasks. The data for travel toward and travel away have been combined since direction of movement did not affect the duration of travel (see Table 2).

In order to determine which of the magnification conditions differed from each other, data from the two precision conditions were combined and t tests were computed between individual pairs of means (Lindquist, 1953). Travel movements made under $20\times$ were significantly faster than movements made under $30\times$ ($p < .05$) and $40\times$ ($p < .01$). The $30\times$ and $40\times$ conditions did not differ significantly.

DISCUSSION

Results of Experiment I support earlier findings (Nayyar & Simon, 1963) that $30\times$ magnification is superior to $20\times$ for picking up .010-inch diameter dots. Perhaps the most striking finding, however, was the significant interaction between magnification and object size; that is, the optimum magnification varied depending upon the size of the object manipulated. While the higher power ($30\times$) was superior for picking up the smaller dots (.010 inch), a lower power ($20\times$) was superior for picking up the larger dots (.015 inch). Results also indicated that if one magnification had to be used to pick up both size dots, $20\times$ would be the better of the two. Pickup time was not affected by the size of the object manipulated, at least within the range of sizes investigated in this study.

If one may be permitted to speculate on the basis of these very limited data, it would appear that the optimum magnification for pickup is related to the apparent size of the

TABLE 2

SUMMARY OF ANALYSIS OF VARIANCE:
TRAVEL TIME

Source	df	MS	F
Between Ss	23		
P	1	5.7041	22.28**
M \times O _b	2	.1920	—
P \times M \times O _b	2	.8654	3.38
Error _b	18	.2560	
Within Ss	120		
M	2	.1634	4.30*
O	2	.5137	13.52**
M \times O _w	2	.0268	—
P \times M	2	.0439	1.16
P \times O	2	.0142	—
P \times M \times O _w	2	.0054	—
Error _{1w}	36	.0380	
M \times D	2	.0014	—
O \times D	2	.0274	—
M \times O _w \times D	2	.0128	—
P \times M \times D	2	.0094	—
P \times O \times D	2	.0123	—
P \times M \times O _w \times D	2	.0020	—
Error _{2w}	36	.0869	
D	1	.0058	—
M \times O _b \times D	2	.0430	—
P \times D	1	.1682	3.14
P \times M \times O _b \times D	2	.0424	—
Error _{3w}	18	.0535	
Total	143		

Note.—P = Precision, M = Magnification, O = Order,
D = Direction of travel.

* $p < .05$.

** $p < .001$.

object when viewed through the microscope. If apparent size is operationally defined as the product of magnification and object size, it will be noted that the best performance in this experiment was achieved on the two conditions where the apparent size of the dot was .3 inch ($20 \times .015 \text{ inch} = .3 \text{ inch}$; $30 \times .010 \text{ inch} = .3 \text{ inch}$). The results suggest that the apparent size of .3 inch may be a useful guide for selecting the magnification-object size combination which would be most compatible to the worker performing subminiature assembly operations. This suggestion can be validated, however, only by expanding the scope of the present experiment to investigate a wider range of magnifications and object sizes.

Results of Experiment II confirm earlier indications (Nayyar & Simon, 1963) concerning the effects of magnification on travel movements. It appears that increasing the magnification results in a slowing of movements through the magnified field. Thus, while a high power may reduce the time required to pick up a very small object, this same high power may increase the duration of the subsequent transport movement, assuming that it, too, is performed under magnification.

The difference in duration of travel between the two precision conditions in Experi-

ment II also agrees remarkably well with earlier findings. Nayyar and Simon (1963) used a dot loading operation which consisted of assembling .010-inch dots into different diameter holes. They defined precision in terms of the ratio of hole size to object size; for example, in Precision Level 2, the hole was twice the size of the dot, while in Precision Level 5, the hole was five times the size of the dot. They found a 38% slowing in the travel loaded movement under the higher precision condition. In Experiment II, which involved a wire positioning operation performed under two comparable precision conditions, there was a 34% slowing in the travel movements under the higher precision condition. It should also be emphasized that in neither of these experiments was there any evidence of an interaction between precision and magnification. This is in contrast to the marked interaction of object size and magnification noted in Experiment I.

REFERENCES

- LINDQUIST, E. F. *Design and analysis of experiments in psychology and education*. Boston: Houghton Mifflin, 1953.
- NAYYAR, R. M., & SIMON, J. R. Effects of magnification on a subminiature assembly operation. *J. appl. Psychol.*, 1963, 47, 190-195.

(Received January 7, 1963)

SPEED OF READING IN AN ADULT POPULATION UNDER DIFFERENTIAL CONDITIONS¹

ALAN M. KERSHNER

United States Air Force Systems Command, Bedford, Massachusetts

420 adults read 2 pairs of passages and unknown to them each reading time was recorded. After reading the 1st pair of passages, readers were asked to answer a comprehension question. Irrespective of the difficulty level of the passages, reading times for the pair of passages following the comprehension question were significantly longer than for that preceding it. Substantial positive correlations were found between 1st- and 2nd-pair reading times for both individuals and for passages. Slow readers showed greater increases in reading time for more difficult materials than did fast readers. Evidence of the dubious value of words per minute as a measure of reading speed is presented.

Speed of reading, typically, has been measured on individuals within school or other relatively homogeneous populations who knew they were being timed. The primary purpose of this report is to present reading speeds for a general adult sample that was *not* aware of being timed. The whole subject may be of more than ordinary interest in view of the seemingly apocryphal reading rates being reported in the popular press where speeds from 14,000 to 25,000 words per minute are cited (Anonymous, 1960; Kelly, 1962). The concern here is not with how fast people can read, but, rather (a) with how fast people *did* read under two differing conditions; (b) with the generality of the effects for reading materials of differing levels of difficulty; (c) with interactions between materials of differential difficulty and readers of differential skill; and (d) with questioning the conventionally used "measure" of words per minute.

Four hundred and twenty adults, with an estimated age range from 18 to 85 and a reported education from Grade 4 through 5 years of graduate study, each read four *different* passages. The subjects, 147 males and 273 females, were obtained from 63 randomly selected blocks within the District

of Columbia. Thirty-seven reading passages of widely varying difficulty were presented in a modified paired-comparison arrangement; the 37 passages were divided into 2 main groups of 21 each with 5 of the passages being common to both groups. Each passage was restricted in length to one side of an 8.5×11 inch page of double-spaced elite type. Four concealed stop watches were used to secure the reading times; this procedure was presented fully by Hackman and Kershner (1951).

The task of each subject was to judge which passage within each of two pairs of passages was the more difficult for him to read. After judging which of the first two passages was the more difficult for him to read and giving his reasons therefor, each subject was presented with a comprehension question. During the reading of the first pair of passages, subjects lacked knowledge that a comprehension question was to be asked, whereas, it may be assumed, during the reading of the second pair of passages, they expected a comprehension question. Accordingly, it is believed that situations of differential stress resulted.

1. Figure 1 presents two distributions of the time in seconds taken to read passages as members of the first pair as contrasted with the time taken to read passages as members of the second pair. Because of the variation in length of individual passages, the reading time for each passage was equated to 2,000 type spaces. All reading times included in this

¹ This is ESD-TDR-62-192 of the Air Force Electronic Systems Division. The analysis presented here was done in part under AFSC Project 9674. The primary data were collected under Nonr-01700, ONR Project 153-024 under contract between the Office of Naval Research and the University of Maryland (Hackman & Kershner, 1951).

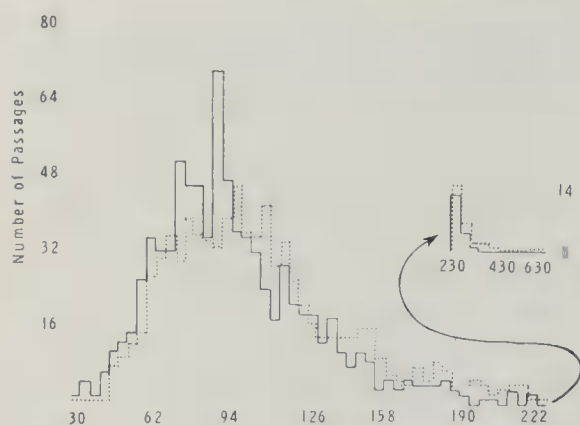


FIG. 1. First- and second-pair reading times in seconds per 2,000 type spaces for 420 adults.

report have been so derived from the original initial reading times.

The range of reading times for passages read as members of the first pair was from 30 to 365 seconds whereas for passages read as members of the second pair it was from 30 to 728 seconds. The median for first-pair passages was 93 seconds as contrasted with 104 seconds for second-pair passages. The mean reading time and standard deviation for first-pair passages were 101.6 and 40.5 seconds, respectively, while for second-pair passages they were 115.4 and 56.6 seconds. The mean reading time per 2,000 type spaces for all passages was 108.5 seconds.

Since each passage was read an equal number of times in each of the four positions, summary data for each position may be of interest. Passages read as members of the first pair had almost identical mean reading times (102.4 versus 100.9 seconds) and standard deviations (40.7 versus 40.2 seconds), respectively, when read in the first and second positions. Within the second pair the mean reading time for passages read in the third position was 118.2 seconds as contrasted with 112.6 seconds for passages read in the fourth position; the standard deviations were 60.0 and 54.0, respectively. Between passages read in third and fourth positions, the ratio of the mean of the differences (5.595) to its standard error (1.942) gives a t value of 2.88 and $.01 > p > .001$ with 419 df . Subjects, while reading the third passage, may have expected they would have to remember content after reading the fourth passage and,

accordingly, attempted to read with special care.

The within-individual differences between first- and second-pair reading times appear to be considerable. For an appropriate test of their significance, differences were computed between the *total* time taken by each person to read the first pair of passages and the *total* time taken to read the second pair of passages. The mean of these differences was 27.56 seconds with a standard error of 3.57. With 419 df , this gives a t value of 7.72, which is more than double the value to reflect significance at the .001 level of confidence (Johnson, 1949, p. 78).

Because of the widely varying difficulty of the various passages used in this experiment, an expression of the relationship found between the total time taken to read each of the two pairs probably underestimates the true correlation between the first- and second-pair reading times. None the less, a fairly high relationship was found—the Pearson r was .72 with a standard error of .02. The correlation between first- and second-pair reading times for 90 individuals whose first- and second-pair passages could be equated for readability was $.81 \pm .04$.

2. Whereas the preceding has dealt with different passages read by the same individuals, this section will consider effects and relationships obtaining when the same passages were read by different individuals; more specifically, with the generality of the effects for reading materials of differing levels of difficulty. An operational definition of readability used in the basic experiment employed the dual criteria of relative reading time and judged difficulty; this permitted the classification of 11 of the reading passages as Hard, 11 as Easy, 7 as Intermediate, and 8 as Inconsistent with respect to the two criteria (Hackman & Kershner, 1951). It may now be asked whether significant differences occurred between first- and second-pair reading times for different classes of reading materials.

Table 1 presents the mean reading times, t , and p values. The t values were computed from the mean reading times for each passage when read as a member of the first or second

pair; 32 means for each pair were based on readings by 20 individuals and 5 of the means were based on 40 readings in each pair. The reader should recognize that Table 1 is presented to show relationships on the horizontal. However, when the table is reconstructed using judged difficulty as the sole classification criterion, the only main effect is to reduce, somewhat, differences in means on the vertical; the order of relationships either horizontal or vertical is not altered, and differences on the horizontal are almost identical.

Although the difference in the mean reading time for the Easy passages is less than half as large as differences for the other classes of reading material, it is, nevertheless, significant beyond the 5% level of confidence.

A determination of the relationship between the mean times taken to read each of the 37 passages as a member of the first versus the second pair yielded a Pearson *r* of .66 with a standard error of .09. Such a substantial relationship between first- and second-pair mean reading times for the same passage when read by different individuals strongly attests to the value of a readability concept.

3. Pertinent to everyday communications problems faced by educational, industrial, business, and governmental organizations is the interaction between readability of different classes of materials and the reading ability of individuals. To shed some light on this, mean reading times were computed for the three readability classifications of Easy,

TABLE 1

MEAN READING TIMES PER 2,000 TYPE SPACES, *t* AND *p* VALUES FOR PASSAGES CLASSIFIED AS TO READABILITY AND READ AS FIRST- OR SECOND-PAIR MEMBERS

Passage classification	Number of passages	First-pair mean	Second-pair mean	<i>t</i>
Hard	11	113.5	129.4	3.013**
Intermediate	7	100.6	116.7	4.050***
Easy	11	90.8	97.9	2.277*
Inconsistent	8	102.3	120.8	5.169***

* *p* < .05.
** *p* < .02.
*** *p* < .01.

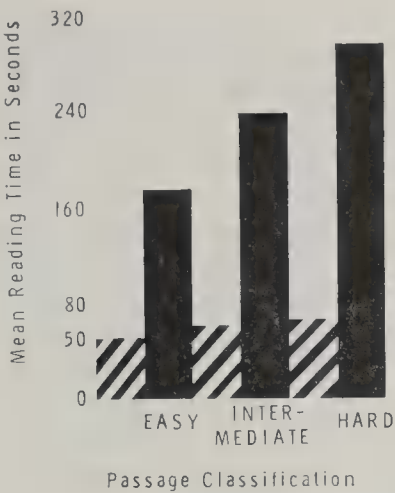


FIG. 2. Mean reading times for 21 fastest and 21 slowest readers for three readability classifications (striped blocks = fastest readers; solid blocks = slowest readers).

Intermediate, and Hard, for passages read by the 21 subjects taking the largest total time and the 21 subjects taking the smallest total time to read the four passages presented. Although the slowest readers did read more Hard (21 versus 15) and less Easy (19 versus 36) passages than the fastest readers, differences for the 3 × 3 table, which included the Intermediate passages, failed to yield a chi square significant at the 5% level of confidence. Figure 2 presents the mean times taken by the slowest and fastest readers for each of the three readability classifications.²

The data indicate that as one ascends a readability difficulty continuum the effect on reading time for slow readers increases differentially over the effect for fast readers. Figure 2 is intended to emphasize the greater increases in reading time of the slow readers as compared to increases among the fast readers. Although increases in both groups may be expected, since passages were classified partially on the basis of reading time, the order of these relationships is not altered when the single criterion of judged difficulty is used as a basis of classification.

4. For those interested in words per minute for a general adult population sample, passages read as members of the first pair were read at an average rate of 201 words per

² Thanks are due to Raymond S. Nickerson for suggesting this look at the data.

minute, whereas second-pair passages were read at 177 words per minute. Because of the task which was posed, it is believed that rates for the first pair may more closely approximate a so-called normal reading rate. The above rates may be compared with other rates of 250 words per minute, estimated as normal for high school students and adults (Harris, 1961), and rates of 280–340 words per minute for college students found by several investigators (Educational Developmental Laboratories, undated).

It should be evident, however, that without knowledge of the reading difficulty of given material, speed of reading data may be of questionable significance. Furthermore, the conventional practice of reporting speed of reading as words per minute can compound the uncertainty. In the research reported here the average number of words per 2,000 type spaces for the 37 reading passages was 340.5. One reading passage taken from the *Annals of the American Academy of Political Science* had only 298 words per 2,000 type spaces, whereas another passage taken from *True Confessions* had 396; these were the extremes of the range. When the overall mean reading time for each passage was correlated with

words per 2,000 type spaces for each passage a relationship of $-.30 \pm .15$ was found. The time is overdue for the use of a less elastic yardstick than words per minute in the reporting of reading speeds.

5. In general, the analysis of these data indicates that reading time alone may serve quite adequately as a criterion of readability. Certainly, further research is warranted to establish both the utility and generality of this criterion for a wider range of materials that should include other types of formats.

REFERENCES

- ANONYMOUS. Read faster and better. *Time*, (Aug. 22), 76, 41–42.
- EDUCATIONAL DEVELOPMENTAL LABORATORIES. *EDL News Letter No. 13*. Huntington, N. Y.: EDL, undated.
- HACKMAN, R. C., & KERSHNER, A. M. The determination of criteria of readability. Technical Report, 1951, University of Maryland, Contract NR 153-024, Office of Naval Research.
- HARRIS, A. J. *How to increase reading ability*. New York: Longmans, Green, 1961.
- JOHNSON, P. O. *Statistical methods in research*. New York: Prentice-Hall, 1949.
- KELLY, T. Speedy reading. *Wash. Daily News*, 1962 (May 31), 20.

(Received January 14, 1963)

A NOTE ON THE NAGLE ATTITUDE SCALE

ROBERT M. GUION AND JAMES E. ROBINS¹

Bowling Green State University

Nagle's scale to measure attitude toward supervisor assumed a relationship between an attitude toward the supervisor and a perception of how he behaves. This assumption was rejected by Brayfield and Crockett (1955). Evidence is now presented in support of the assumption and in rebuttal to Brayfield and Crockett. A special scale to measure attitudes toward supervisors was constructed specifically to avoid items referring to overt behavior. The correlation between the 2 scales for 74 research and engineering employees was .786, or .835 when corrected for attenuation.

In an effort to study the relationship between employees' attitudes and their productivity, Lawshe and Nagle (1954) used a questionnaire developed by Nagle (1953) as a measure of attitude toward one's immediate supervisor. Items were phrased with reference to supervisory behavior; employees were asked to indicate whether their supervisors gave straight answers to their questions, avoided employees who wanted to see them, followed through on promises, etc. The authors assumed that an employee's perception of the behavior of the boss is functionally related to his attitude toward the boss.

Using this questionnaire, they found a correlation of .86 (significant at the 10% level of confidence) between paired-comparison productivity ratings and the means of the attitudes toward the supervisor in 14 work groups. They were quite cautious in making interpretations, pointing out particularly their uncertainty about meaning of the productivity ratings. Nevertheless, they considered it a reasonable and conservative inference that employees of more productive departments viewed their supervisors more favorably.

This conclusion was rejected by Brayfield and Crockett (1955). Their survey of the literature on employee attitudes concluded that there was no relationship between attitude and productivity. The Lawshe-Nagle findings were dismissed as irrelevant to this problem; that is, since the items were descriptions of behavior, it was argued, the results described only a relationship between what a

supervisor actually does and the productivity of his work group.

The question at issue is whether employees can perceive supervisory behavior and report on these perceptions to a degree reasonably independent of their attitudes toward the person who behaves. If they can, then Brayfield and Crockett are perhaps making a reasonable case for excluding the Lawshe-Nagle findings from consideration of the attitude-productivity hypothesis. If not, however, then the Lawshe-Nagle conclusions on this hypothesis (with, of course, the reservations they mentioned) may be considered support for the idea that attitude can be related to productivity. It would seem that responses to such a questionnaire, and to a more traditional attitude scale as well, would necessarily be influenced both by the objective behavior of the supervisor and by the respondent's attitudes toward him.

The Brayfield and Crockett position is based on examination of item content. This is by no means an adequate basis for determining the dimensions underlying a set of "scores" on an instrument such as the Nagle questionnaire, or for determining whether the larger part of the reliable variance is due to objective description of behavior or to the attitudes of the describer. Nevertheless, the distinction seen by Brayfield and Crockett on examination of content needs consideration.

If a questionnaire can be discussed as not measuring attitudes because its items call for descriptions of supervisory behavior, then, by the same a priori logic, an attitude scale consisting of items of affective reaction cannot be said to be descriptive. In other words, the

¹ This study was completed in connection with his thesis research.

Brayfield and Crockett position logically suggests the hypothesis of little or no correlation between the scores derived from essentially descriptive statements and those derived from statements which stress affect rather than behavior.

Such a series of statements was developed by Robins (1961) in a study attempting to partial out the effect of the affective component of descriptive responses. For this he developed an Attitude toward Supervisor scale which minimized descriptive content. The items were taken from rating scale data published for 724 items scaled by the equal-appearing intervals method (Uhrbrock, 1950). A set of 46 items was selected from this pool which was: relatively unambiguous; spread throughout the scale range; and most important, judged to be essentially affective in content rather than behavioral.

To illustrate the latter judgment, the statement, "Gets production out in less time than average," was rejected as referring to a rather objective and specific aspect of performance. On the other hand, a statement with a quite similar mean and standard deviation, "Is very popular with fellow employees," was deemed acceptable as essentially affective and much more difficult to pin down behaviorally. A few items did contain behavior reference, but they were included only if the content appeared affective by stressing an evaluation of behavior; for example, "Delegates work to others wisely."

Obviously, such a set of items can no more succeed in completely eliminating description of behavior than does the Nagle questionnaire succeed, from the Brayfield and Crockett argument, in completely eliminating attitude. Nevertheless, if the Lawshe-Nagle results is justified, the correlation between these scales should be low.

As a matter of fact, the correlation was .786, or .835 when corrected for attenuation in both variables, for a sample of 74 salaried employees in research and engineering departments. In short, 70% of the reliable variance

in the Nagle questionnaire, with items written as behavioral descriptions, is common with variance of the Robins scale, in which behavioral objectivity of item content was deliberately avoided.

It is not suggested that one can identify sources of variance by argument; rather, the intent of this note is to offer evidence that affect and perception of behavior are so thoroughly mixed in questionnaires of either sort that the Nagle questionnaire cannot be so easily dismissed from attitude measurement. As a matter of fact, there may well be merit in devising for attitude measurement questionnaires which give the impression of objectivity.

Although it is useful to clarify earlier research, it is also desirable to consider future problems. A reasonable question in planning future studies would concern the relative merits of the Nagle questionnaire and the scale used here as a criterion. In many situations in which it is desired to know something of the employees' attitudes toward their direct supervisor, it is also considered undesirable to be too blatant about it. Of the two measures used in this study, it would appear that the Nagle questionnaire might be more desirable in such situations because of its apparent concern with objectivity as an inventory of supervisory behavior.

REFERENCES

- BRAYFIELD, A. H., & CROCKETT, W. H. Employee attitudes and employee performance. *Psychol. Bull.*, 1955, 52, 396-424.
- LAWESHE, C. H., & NAGLE, B. F. Productivity and attitude toward supervisor. *J. appl. Psychol.*, 1954, 37, 159-162.
- NAGLE, B. F. Productivity, employee attitude, and supervisor sensitivity. Unpublished doctoral dissertation, Purdue University, 1953.
- ROBINS, J. E. A study of the relationship between job satisfaction and congruence between expected and perceived supervisory behavior. Unpublished master's thesis, Bowling Green State University, 1961.
- UHRBROCK, R. S. Standardization of 724 rating scale statements. *Personnel Psychol.*, 1950, 3, 285-316.

(Received January 21, 1963)

JOB ATTITUDES IN MANAGEMENT:

V. PERCEPTIONS OF THE IMPORTANCE OF CERTAIN PERSONALITY TRAITS AS A FUNCTION OF JOB LEVEL¹

LYMAN W. PORTER AND MILDRED M. HENRY

University of California, Berkeley

This study investigated managers' perceptions of the relative importance of 10 personality traits for success in their managerial roles, as a function of level of position within management. Data were obtained by means of a questionnaire in which the 10 traits were rank ordered in importance by over 1800 respondents from all parts of management and all types of companies. The 10 traits consisted of 5 Other-Directed or Organization Man type traits, and 5 Inner-Directed type traits. Results showed that Inner-Directed traits were perceived as more important at each higher level of management and thereby Other-Directed traits were seen as more important at each lower level of management.

Earlier articles in this series on job attitudes in management have dealt with perceptions of the importance and degree of satisfaction of certain psychological needs relevant to the work situation. The present study examines a different type of dependent attitude variable, namely, perceptions of various personality traits in terms of their relative importance for "success" in the managerial job. However, this study is not intended as merely another attempt to apply the "trait approach" to the problem of organizational leadership. It is instead aimed at trying to find out in more detail how managers think about and look at their positions within the modern business organization. Thus, the focus of the present study is not on the actual rankings of the traits as such, but rather on what these rankings reveal concerning how individuals at different levels of the organizational hierarchy view the psychological nature of their managerial and executive roles.

A recent paper on this general topic (Porter, 1961) examined the relative importance attached to 13 different personality

traits at two levels of management: the bottom level and the next level above. Those 13 particular traits were selected for study to represent some of the more important dimensions measured in widely used personality tests. Results from that earlier study showed that the two levels of management tended to be in agreement in how they ranked each trait with regard to its importance for managerial success. However, there also was some evidence to suggest that a certain type of trait increased somewhat in importance from the bottom to the next higher management level, and an opposite type of trait decreased in importance between these two levels. The group of items which was regarded as relatively more important at the higher of the two levels was composed of the traits "aggressive," "dominant," "independent," and "original"; while the group which decreased in importance included "conforming," "co-operative," "flexible," and "sociable." Since the groupings of these two clusters of traits were carried out entirely post hoc rather than prior to the study, no statistical treatments were applied to them and no formal names were attached to them. Nevertheless, it appeared that the first group of items represented traits "showing a strong emphasis on personal and individual capabilities," while the second group contained traits showing a "concern for adapting to the feelings and behavior of others" (Porter, 1961).

The results of the previous study can only

¹ This study was carried out as part of the research program of the Institute of Industrial Relations, University of California, Berkeley. The data were collected while the senior author was a Ford Foundation Faculty Research Fellow. The Institute of Social Sciences at the University of California and the American Management Association contributed to the support of the research assistance, and the Computer Center of the University provided facilities for data computation.

be regarded as suggestive, as the study was not designed specifically to investigate the relative perceived importance of these two different types of traits in different parts of management. Therefore, the present study was undertaken to provide evidence of a more systematic nature regarding changes in the perception of the importance of certain types of personality traits from lower to higher management. Especially, the present investigation was aimed at providing a more direct comparison of relative differences in importance between the two types of personality traits described above among different management levels. To accomplish this objective, two clusters of traits were set up to represent the two ends of the dimension that seemed to exist in the previous set of findings. However, the specific traits chosen for the new study were for the most part not the identical ones that appeared in the earlier article; instead, new traits were selected in order to focus more directly on the two types of traits suggested by the previous evidence. The personality trait dimension to be investigated here is intended to be a composite dimension made up of related but slightly different personality continua. One of these specific continua is Riesman's Inner-Directed-Other-Directed distinction (Riesman, 1950). Another is Whyte's (1956) description in *The Organization Man* of the Protestant Ethic versus the Social Ethic as behavior guides in management. A third source of the composite dimension was the contrasting picture of self-descriptions of top managers versus middle managers in a previous article on self-perceptions (Porter & Ghiselli, 1957). Still another related continuum is one which has been termed the "entrepreneurial-bureaucratic" dimension. Considered together, then, these various specific continua listed above have much in common, but each is slightly divergent in emphasis from the others. The two clusters of traits used in the present study constitute an attempt to include aspects of each continuum in a composite dimension. The terms "Inner-Directed" and "Other-Directed" were chosen as labels for the two ends of the composite dimension because they immediately connote the nature of the dimension and because Riesman's description

of these character types was a major basis for the selection of the specific traits.

The problem investigated in this study was the following: Does the perception of personality qualities needed for managerial job success change from lower to higher levels of management? More specifically, are Inner-Directed (Protestant Ethic, Top Management Self-Description) traits regarded as increasingly more important for success at each higher level of management relative to Other-Directed (Social Ethic, Middle Management Self-Description) traits?

METHOD

Questionnaire

Data were collected for this study by means of a three-section questionnaire. Two of the sections concerned topics not relevant to the present investigation and will not be described here. The section of the questionnaire pertinent to this study contained the following instructions (in part):

The purpose of [this part of the questionnaire] is to obtain a picture of the traits you believe are most necessary for success in YOUR PRESENT MANAGEMENT POSITION.

Below is a list of 12 traits arranged randomly. Rank these 12 traits from 1 to 12 in the order of their importance for success in your present management position.

While the respondents were asked to rank 12 personality traits, 2 of the traits—Intelligent and Efficient—were camouflage items put into the list to disguise the dimension being studied. In analyzing the results, the ranks for these 2 items were not counted and were removed from the list, and the remaining 10 items were reranked from 1-10. Thus, for example, if Intelligent had been ranked third and Efficient sixth, by a given respondent, the traits ranked fourth and fifth were moved up one notch in rank, while the traits originally ranked from seventh to twelfth were moved up two notches. In this way, the 2 camouflage items were prevented from having any direct effect on the ranks of the 10 relevant traits.

The 2 camouflage traits and the 10 relevant traits were presented in a random order in the questionnaire. The 10 relevant traits are listed below in the two theoretical clusters used as the basis for the analysis of the results.

Inner-Directed Cluster	Other-Directed Cluster
Forceful	Cooperative
Imaginative	Adaptable
Independent	Cautious
Self-Confident	Agreeable
Decisive	Tactful

Procedure and Sample

The procedure for obtaining the sample of respondents has been described in detail in a previous paper (Porter, 1962). The questionnaire was sent by mail to approximately 6,000 managers in all types of companies located throughout the country.² Responses were received from 1,958 individuals for a response rate of 33%. Of these 1,958 returns, 1,896 had the relevant portion of the questionnaire filled in correctly and constitute the sample for this study. Approximately two-thirds of the sample were managers from manufacturing companies, while the other third represented nonmanufacturing firms.

Answers to personal data questions on the questionnaire permitted the classification of respondents according to a number of independent variables. Two of the variables are relevant for this study: level of position within management and age of the respondent. The major independent variable under investigation here is level in the management hierarchy. Consequently, a five-category system of classification on this variable was set up. These categories were (from top to bottom):

President:	presidents and chairmen of boards
Vice President:	vice presidents (or their equivalents in large companies)
Upper-middle:	approximately the level of division managers, plant managers, and major department managers
Lower-middle:	approximately the level of department and subdepartment managers
Lower:	first-level or second-level supervisors

The method of assigning each respondent to one of the five level-of-position categories listed above has been described elsewhere (Porter, 1962).

In order to carry out statistical tests of trends for management levels, respondents were also cross-classified by age as well as by level of position. This was done for two reasons; first, since the age variable is correlated somewhat with the level variable, it was felt necessary to hold age constant in order to assess the effects of level of position. Secondly, by having several categories of age for each management level, it was possible to investigate effects of management level on four independent subsamples of respondents. Consequently, subjects were placed into one of four age groups (as well as into one of five level-of-position groups): 20-34, 35-44, 45-54, and 55+.

One other variable, amount of formal schooling, was also calculated for the present sample of respondents. However, it was found that the percentage of respondents with college degrees was almost identical (approximately 75%) for each

of the five management levels. Therefore, the results obtained for management levels cannot also be attributed to differences in amount of formal education.

RESULTS

Tables 1 and 2 present results in terms of each of the 10 individual traits ranked by respondents, and they also emphasize the comparative trends for the two clusters of traits: the Inner-Directed cluster and the Other-Directed cluster.

Table 1 shows the mean ranks for each of the 10 traits and for each cluster of traits for each of the five levels of management. The values in each cell in Table 1 for each trait have been subtracted from the constant number 10, so that larger numbers in Table 1 represent greater perceived importance.

Certain trends can be seen in Table 1. First, the cluster scores show that the Inner-Directed cluster of traits was regarded as more important the higher the level of management, while the reverse was true for the Other-Directed cluster. (It should be pointed out that if a certain trend exists for one of the two clusters, the opposite trend must exist for the other cluster, because the results are in terms of relative ranks rather than absolute ratings.) Secondly, in terms of the 10 individual traits, almost all of them show trends in line with the overall trends for their respective clusters. However, it can also be seen that the trends do not hold equally strongly for each trait within a cluster. For example, trends for Self-Confident in the Inner-Directed cluster and Cautious in the Other-Directed cluster were weak or non-existent.

The trends evident in Table 1 are shown with greater clarity in Table 2, where levels of statistical significance can be attached to the trends. Table 2 is based on values obtained by subdividing the sample by age as well as by management level.³ Thus, the

³ The table presenting these mean values for each management level within each age group has been deposited with the American Documentation Institute. Order Document No. 7688 from ADI Auxiliary Publications Project, Photoduplication Service, Library of Congress, Washington, D. C. 20540. Remit in advance \$1.25 for microfilm or \$1.25 for photocopies and make checks payable to: Chief, Photoduplication Service, Library of Congress.

² The assistance of the American Management Association, and particularly Robert F. Steadman, in obtaining the sample of respondents is gratefully acknowledged.

effects of age are held constant and four separate samples of managers are created to study the effects of management level independent of age. Table 2 is designed to show the changes in mean importance of each trait from higher to lower levels of management within each of the four age groups. Under each age group column in Table 2 are sub-columns labeled +, 0, and -. The numbers below each of these three columns within each group represent the number of increases, decreases, and no changes, respectively, in the mean importance of a trait from the upper to the lower levels of management. Whenever a mean increased by more than .05 of a mean rank unit from one management level to the next lower level, it was counted as an increase or +. Whenever the mean decreased by more than .05 of a mean rank unit, it was counted as a decrease or -. If the change was .05 or less of such a unit, it was counted as no change or 0. For example, for the trait Forceful: in the 20-34 age group its mean rank increased once and decreased twice in going from the Vice President to the Lower management level. In the 35-44 age group

Forceful increased in importance once and decreased three times from the President to the Lower level. In the other two age groups Forceful increased once and decreased five times in importance from higher to lower management levels. Adding across all four age groups shows that for Forceful there were three increases and 10 decreases between upper and lower management. Clearly, then, this particular trait was considered relatively more important for success at higher levels of management compared with lower levels.

The results presented in Table 2 show the following: (a) All five Inner-Directed traits, except Self-Confident, decreased in importance from higher to lower levels of management. Self-Confident showed no trend to either increase or decrease in importance. (b) Four of the five Other-Directed traits were seen as more important for success at lower levels of management compared with upper levels. The one exception was Cautious, which tended to be seen as more important at higher than at lower management levels. (c) The trends for the majority of Inner-Directed and Other-Directed traits held strongly for all

TABLE 1
MEAN IMPORTANCE (10- \bar{X} rank) OF TRAITS FOR JOB SUCCESS BY MANAGEMENT LEVEL

Trait	Management level				
	President (<i>N</i> = 112)	Vice President (<i>N</i> = 604)	Upper- Middle (<i>N</i> = 650)	Lower- Middle (<i>N</i> = 428)	Lower (<i>N</i> = 102)
Inner-Directed					
Forceful	4.66	4.19	4.02	3.82	3.37
Imaginative	7.37	6.94	6.73	6.41	6.44
Independent	2.96	2.70	2.54	2.36	2.44
Self-Confident	5.86	5.43	5.36	5.87	5.53
Decisive	6.71	6.08	5.96	5.61	5.35
Total for cluster	27.56	25.34	24.61	24.07	23.13
Other-Directed					
Cooperative	4.54	5.49	5.61	5.88	5.75
Adaptable	4.54	4.77	5.13	5.16	5.13
Cautious	1.30	1.38	1.30	1.24	1.47
Agreeable	2.18	2.61	2.84	2.91	3.89
Tactful	4.89	5.42	5.52	5.74	5.63
Total for cluster	17.45	19.67	20.40	20.93	21.87

Note.—Higher numbers indicate greater importance.

TABLE 2

CHANGES IN MEAN IMPORTANCE OF TRAITS FOR JOB SUCCESS FROM HIGHER TO LOWER LEVELS OF MANAGEMENT WITHIN AGE GROUPS

Trait	Age group												Total group		
	20-34			35-44			45-54			55+					
	+	0	-	+	0	-	+	0	-	+	0	-	+	0	-
Inner-Directed															
Forceful	1	0	2	1	0	3	0	0	3	1	0	2	3	0	10 ^a
Imaginative	0	0	3	1	0	3	0	1	2	2	1	0	3	2	8
Independent	2	0	1	1	0	3	0	0	3	1	0	2	4	0	9
Self-Confident	2	0	1	1	0	3	2	0	1	1	1	1	6	1	6
Decisive	1	0	2	0	1	3	1	0	2	0	1	2	2	2	9 ^a
Total	6	0	9	4	1	15	3	1	11	5	3	7	18	5	42*
Other-Directed															
Cooperative	1	1	1	3	0	1	3	0	0	2	0	1	9	1	3
Adaptable	1	0	2	2	1	1	1	0	2	3	0	0	7	1	5
Cautious	1	0	2	1	2	1	0	2	1	1	0	2	3	4	0
Agreeable	2	1	0	3	0	1	3	0	0	3	0	0	11	1	1
Tactful	3	0	0	2	2	0	2	1	0	2	0	1	9	3	1
Total	8	2	5	11	5	4	9	3	3	11	0	4	39	10	16*

Note.—A sign test was used to determine *p* values.
^a Approaches significance (*p* = .10).
^{*} *p* < .01.

age groups except the 20-34 group. In this latter age group the trends held only slightly. Summarizing the results shown in Table 2: There were statistically significant trends for Inner-Directed type traits to be regarded as more important and Other-Directed type traits less important at higher compared with lower levels of management.

DISCUSSION

The results of this study point clearly to the fact that the psychological demands of the job, in terms of relative emphases on different types of personality qualities, change from one part of the management structure to another part. At the lower levels, apparently, a manager to be successful in his position should exhibit aspects of both Inner- and Other-Directed behavior. The same conclusion also holds for upper levels: both Inner- and Other-Directed behavior is important. Nevertheless, something further needs to be added to these two conclusions: the *relative* emphasis on Inner-Directed behavior increases

at higher levels of management and the emphasis on Other-Directed behavior correspondingly decreases at these levels. Thus it would appear that if there is an Organization Man, he would more likely be found at the lower rather than the upper levels of the organization.

Behind the rather broad generalizations given above are some important qualifications. In examining the effects of management level, it was found that four of the five traits labeled Other-Directed were seen as consistently more important at each lower level of management. This was especially true for two traits central to the Other-Directed concept, Cooperative and Agreeable. The one trait that had been put into the Other-Directed category that did not show increasing importance with lower levels of management was the trait of Cautious, which was regarded as slightly more important at higher than at lower levels. As a personality trait, Cautious can be regarded as somewhat less relevant to the Riesman-type Other-Direction

concept than are the other four traits in that category in this study. However, Cautious does seem central to Whyte's picture of the Organization Man. The results for the five Inner-Directed traits showed that four of them consistently were seen as more important at higher levels of management. The trend was especially strong for two key Inner-Directed type traits, Forceful and Decisive. The one Inner-Directed trait that showed no trend for greater or lesser importance at higher levels was Self-Confident. The reasons why this trait failed to increase in importance at upper levels of the hierarchy are not clear. Theoretically it should have, but empirically it did not. The safest conclusion seems to be that managers see greater general necessity for Other-Directed, Social Ethic, Organization Man behavior at lower managerial levels, but this greater necessity does not extend to

all specific aspects of this overall dimension. The results therefore suggest some caution in accepting overgeneralized stereotypes and conclusions about the presence of this type of behavior at particular levels of management.

REFERENCES

- PORTER, L. W. Perceived trait requirements in bottom and middle management jobs. *J. appl. Psychol.*, 1961, **45**, 232-236.
- PORTER, L. W. Job attitudes in management: I. Perceived deficiencies in need fulfillment as a function of job level. *J. appl. Psychol.*, 1962, **46**, 375-384.
- PORTER, L. W., & GHISELLI, E. E. The self perceptions of top and middle management personnel. *Personnel Psychol.*, 1957, **10**, 397-406.
- RIESMAN, D. *The lonely crowd*. New Haven: Yale Univer. Press, 1950.
- WHYTE, W. H., JR. *The organization man*. New York: Simon & Schuster, 1956.

(Received January 21, 1963)

GRAPHIC RATINGS AND ATTITUDE MEASUREMENT:

A COMPARISON OF RESEARCH TACTICS¹

JAMES B. TAYLOR AND HOWARD A. PARKER

University of Washington

This paper suggests that under certain conditions attitudes may be measured as validly and as reliably with a single "attitude report question" as with a multi-item attitude scale. 84 physicians responded to 8 Guttman scales, and also rated their attitudes on 8 comparable graphic rating scales. These data were factor analyzed, and the communalities used as an estimate of the minimal scale reliability. In general, the graphic rating scales proved as reliable as the Guttman scales. An examination of the interscale correlations showed that similar conclusions would be drawn from either technique. It is suggested that a single graphic rating may usefully substitute for a multi-item attitude scale when the attitude continuum is unidimensional.

This note deals with a problem which caused some controversy during the 1940s, without being at that time solved. The question is this: can an attitude be measured as validly and reliably with a single question as with a multi-item attitude scale? The common objections to single attitude questions center around the twin problems of reliability and interpretability. It is usually argued that a single item is more apt to be influenced by chance factors than is a score based on many items, so that a single attitude question will necessarily be less reliable. McNemar (1946) has suggested, for instance, that:

the reliability for single questions will be found to be low; if measured by some coefficient equivalent to the product-moment correlation, we estimate that typical reliabilities will be in the .50's or .60's, sometimes lower and occasionally higher [p. 325].

This conclusion was apparently advanced on logical grounds, since the studies he reviewed had conflicting implications: one, for instance, finding essential equivalence in the results given by graphic rating scales and multi-item Thurstone scales (Riker, 1944), another finding low reliability for "equivalent forms" of single attitude questions (Hayes, 1939).

The problem of interpretability has been raised by Eysenck and Crown (1949). In

analyzing single questions *within* an anti-Semitism scale, he found two of them to correlate .91 with an anti-Semitism factor; but pointed out that it would be hard from these questions to arrive at a decision as to who was anti-Semitic, or to judge the proportion of people who held anti-Semitic attitudes.

The present note deals with the reliability and usefulness of a certain type of attitude item, here called an "attitude report question." Such a question may be defined as one which subsumes all other attitude items in the same attitude universe. The difference between an attitude report question and the usual scale item is exemplified in the difference between asking, "Do you feel that Jews corrupt everything with which they come into contact?" (Eysenck & Crown, 1949, p. 75) and asking, "In general, how do you feel about Jews?" with the degree of favorability being rated on a graphic scale. It should be noted that in most cases an attitude report question is no more obvious than the usual attitude scale item, and its results are considerably more interpretable. The attitude report question does have one restriction not present in all multi-item scales: it necessarily assumes that one and only one attitude dimension is being measured. Apart from this, its main problem lies in its reliability and in the comparability of its findings with those obtained from multi-item scales.

The study reported below compares eight

¹ This study was supported in part by a grant from the Western Interstate Commission for Higher Education, Boulder, Colorado, as part of a larger research assessing the effectiveness of postgraduate medical training in psychiatry.

brief Guttman scales with eight "equivalent" attitude report questions. Guttman scales would seem most appropriate to such a comparison, since the Guttman scaling method results in short scales whose items, like the single attitude report question, tap only one attitude dimension.

METHOD

In a study of physician attitudes, eight Guttman scales were constructed. The items used and the attitudes measured were based on earlier interview and questionnaire studies (Ogle & Taylor, 1961; Taylor, 1961). The scalogram analysis used mail-back responses from 84 physicians, members of the Washington Academy of General Practice; scale patterns were estimated with the Goodenough technique after dichotomizing each item.

The same 84 physicians were given eight attitude report questions, developed on an a priori basis to measure the attitude continua sampled by the Guttman scale. No attempt was made to pretest the questions; in effect the General Practitioners were simply asked to rate their attitudes from 1 to 7. The instructions were as follows:

Here we would like your opinion on some general questions concerning the emotionally disturbed patient and Psychiatry. Please circle the number on the seven-point scale that best expresses your opinion.

Immediately thereafter, the respondent found eight attitude report questions, all in the following format:

Questions

How do you feel about treating emotionally disturbed patients?

(Circle one number for each question)

Dislike
Treating

1

7

Like
Treating

RESULTS

Table 1 shows the eight attitude report questions, and the scale statistics for the corresponding Guttman scales (i.e., Reproducibilities, Minimum Marginal Responses, and Number of items).²

² A 2-page table, giving the Guttman scale items for each scale, has been deposited with the American Documentation Institute. Order Document No. 7689 from ADI Auxiliary Publications Project, Photoduplication Service, Library of Congress, Washington, D. C. 20540. Remit in advance \$1.25 for

All scale scores and attitude report questions were intercorrelated. Also included in the correlations were the number of years in practice in the present community, the year in which the MD degree was obtained, the number of patients seen per day, and the physician's judgment of the percentage of his patients with some significant emotional disturbance. The resultant 20 × 20 correlation matrix was factor analyzed using the principal axis method. Five orthogonal factors were extracted; these were not rotated to simple structure since the psychological meaningfulness of the factors was not at issue.

These data, first of all, show the correlations between the Guttman scales and the corresponding attitude report questions. Secondly, by examining the intercorrelations among all scales, one can see whether the use of attitude report questions alone would have led to different conclusions than the use of Guttman scales alone. And, thirdly, by examining the communalities of each scale and each attitude report question, one can estimate the amount of reliable and shared variance in each scale. Of course, the systematic variance of a scale is likely to be greater than that estimated by the communality, since a scale usually measures something apart from the factors it shares with other scales. With the present data however, the shared variance is great enough to suggest that the unique variance is unimportant, and that the communality figures provide a realistic estimate of the lower limits of scale reliability.

The correlations between the attitude report questions and the corresponding Guttman scales are presented in Table 1. Ranging from .70 to .34, the coefficients are generally substantial but not outstandingly high. It is probable that a correction for attenuation would raise them markedly, but without clarifying sources of disagreement or covariance. A clearer picture is given by the principal axis factor analysis of the entire correlation matrix.

microfilm or \$1.25 for photocopies and make checks payable to: Chief, Photoduplication Service, Library of Congress.

TABLE 1

ATTITUDE REPORT QUESTIONS, STATISTICS FOR COMPARABLE GUTTMAN SCALES,
AND CORRELATIONS BETWEEN THE TWO METHODS

Attitude report question	Statistics of comparable Guttman scale			
	Number of items	Rep.	MMR	Attitude report-Guttman scale correlation
1. How do you feel about treating emotionally disturbed patients?	6	.89	.70	.57
2. In general, do you feel that the treatment of emotional problems should be a major part of the G.P.'s job?	5	.86	.66	.65
3. In your opinion, how adequate has been your overall training in dealing with emotionally disturbed patients?	6	.89	.72	.65
4. How satisfied are you with your present treatment of emotionally disturbed patients?	4	.82	.61	.51
5. In general, how do you feel about the results of psychiatric treatment?	5	.84	.62	.70
6. In general, what has been your impression of psychiatrists?	4	.87	.66	.47
7. How adequate is the care given to institutionalized mental patients in your opinion?	4	.80	.63	.47
8. How difficult is it for the G.P. to treat emotional problems, given the usual limitations of time and finances?	5	.85	.70	.34

Note.—Rep. = Reproducibilities; MMR = Minimum Marginal Responses.

It was found that five orthogonal factors accounted for 62% of the total variance of the scales. Factor I has been tentatively labeled "Interest in the Treatment of the Emotionally Disturbed," Factor II labeled as "Attitudes toward Psychiatry," Factor III is called "Pressures of Practice," Factor IV is seemingly an "Age" component, while Factor V is less easily labeled but is perhaps related to a general tendency towards conservatism. Table 2 shows the factor loadings and the communalities for each of the 20 variables.

A comparison of the Guttman scales and the attitude report questions indicates that

loadings for comparable scales are, in general, similar. In seven of the eight comparisons the rating scale seemed the purer, with the dominant factor being the more highly loaded. In other words, in seven of the eight cases the attitude report question was closer to being unidimensional than was the Guttman scale.

As discussed above, the communalities may be used as a minimum estimate of scale reliability, that is, of shared nonerror variance. The square root of each communality listed in Table 2 gives a figure corresponding to the minimum reliability coefficient. Thus, for the Guttman scales, the minimal

reliabilities would be .73, .81, .79, .79, .84, .76, .73, and .74; for the attitude report questions, .87, .77, .81, .75, .90, .82, .78, and .66. It appears, then, that neither method is consistently the more reliable.

We may ask another question of the two techniques: do they lead to comparable results and generate similar conclusions? One way of answering this question is to examine the intercorrelations among scales, looking

for significant correlations and examining the significance levels. Are the significant correlations between Guttman scales the same as those between comparable attitude report questions? And do the Guttman scales correlate with other measures ("Number of patients seen per day," etc.) in the same way as do comparable attitude report questions? A comparison of findings from the two techniques is given in Table 3. In this comparison

TABLE 2
FACTOR LOADINGS FOR COMPARABLE GUTTMAN AND GRAPHIC RATING SCALES
AND FOR FOUR BACKGROUND VARIABLES

Scale	I	II	III	IV	V	<i>h</i> ²
1. Feelings about the emotionally disturbed patient						
Guttman	.71	-.02	-.11	-.12	-.10	.54
Rating	.78	.09	.02	-.36	-.01	.75
2. Feelings about the legitimacy of the counseling role in general practice						
Guttman	.76	-.15	.11	-.19	.00	.65
Rating	.69	.02	-.08	-.20	-.26	.59
3. Perceived adequacy of psychiatric training						
Guttman	.65	-.32	.12	.26	.14	.63
Rating	.73	-.23	.18	.21	.03	.66
4. Feelings about the adequacy of treatment given						
Guttman	.56	-.45	-.20	.13	.22	.62
Rating	.65	-.23	.21	.10	.19	.56
5. Feelings about psychiatry as a discipline						
Guttman	.48	.68	.02	.05	.01	.70
Rating	.39	.79	.15	.01	.10	.81
6. Feelings about psychiatrists as practitioners						
Guttman	.20	.60	-.38	.12	.15	.58
Rating	.32	.75	.02	.11	.03	.68
7. Feelings about psychiatric institutions						
Guttman	-.02	.11	.12	.46	.55	.54
Rating	.11	.05	.15	.45	.61	.61
8. Feelings about the cost and time necessary for counseling						
Guttman	.44	.01	-.41	-.33	.29	.55
Rating	.50	-.40	-.06	.00	.17	.44
Year of MD degree	.30	-.02	-.21	.56	-.66	
Years of practice	-.26	-.01	.24	-.51	.67	
Estimated percentage of patient disturbance	.23	.05	.35	-.29	-.39	
Number of patients seen per day	-.05	.15	.68	.26	-.10	

TABLE 3

COMPARISON OF INTERCORRELATIONS BETWEEN GRAPHIC RATING AND GUTTMAN SCALES

Scale	1 Patients	2 Role	3 Train- ing	4 Treat- ment	5 Psychi- atry	6 Psychi- atrists	7 Institu- tional care	8 Cost and time
1. Feelings re patient	—							
2. Legitimacy of role	xy	—						
3. Adequacy of training	xy	xy	—					
4. Adequacy of treatment	xy	xy	xy	—				
5. Feelings re psychiatry	xy	xy	0	0	—			
6. Feelings re psychiatrists	y	0	0	0	xy	—		
7. Feelings re institutional care	0	0	0	0	0	0	—	
8. Feelings re cost and time	xy	xy	y	x	0	0	0	—
Year of MD degree	x	y	y	0	0	0	0	0
Years of practice	0	y	0	0	0	0	0	0
Estimated percentage of patient disturbance	0	xy	0	0	0	0	0	0
Number patients seen daily	0	0	0	0	0	0	0	0

an x entry indicates a correlation between a Guttman scale and another variable significant at or beyond the .05 level; a y entry indicates a similar correlation involving an attitude report question. Of the 60 comparisons, 53 give comparable results. In two cases only the Guttman scales correlate significantly; while in five cases only the attitude report questions show significance. It thus appears that conclusions drawn from interscale correlations would be much the same, although not identical, whichever method was used.

DISCUSSION

These findings suggest that single attitude report questions may be as reliable as brief Guttman scales, and that they may give "purer" measures of attitude factors. The two techniques also seem to generate similar, although not identical, conclusions. Can one therefore conclude that sophisticated attitude scaling is unnecessary, that—to put it bluntly—the researcher who painstakingly develops a multi-item attitude scale is like a hunter going after duck with a cannon?

If this were so, it would do much to lighten the researcher's task. The attitude report questions have several advantages over the Guttman scales: they are more obviously interpretable, since a neutral point can be ascribed to them; they cover the full range of attitude, while the Guttman items may not; and they give equal-appearing interval scores rather than rank scores as with Guttman scales.

On the other hand, our present findings are at best tentative, being restricted to a few attitude scales with a limited group of respondents. The assumption of unidimensionality imposes another limitation. A single attitude report question is interpretable only if the reported attitude can be placed on a single dimension; if it cannot, the resultant findings are indeterminant. At present, there seems no way to evaluate the unidimensionality of an attitude continuum save by the use of Guttman scaling or factor-analytic techniques.

We would suggest, then, that scalogram analysis is appropriately used to evaluate the unidimensionality of an attitude universe. If

the universe is unidimensional, it may perhaps be measured as reliably—and much more easily—with a single attitude report question. Indeed, it is possible that the attitude report question may prove to be factorially purer—that is, less influenced by extraneous issues—than would a Guttman scale.

REFERENCES

- EYSENCK, H. J., & CROWN, S. An experimental study in opinion-attitude methodology. *Int. J. Opin. Attitude Res.*, 1949, 3, 47–86.
- HAYES, S. P. The inter-relations of political attitudes. *J. soc. Psychol.*, 1939, 10, 359–398, 503–552.
- McNEMAR, Q. Opinion-attitude methodology. *Psychol. Bull.*, 1946, 43, 289–374.
- OGLE, W., & TAYLOR, J. Experience of psychiatrists working with general practitioners caring for discharged mental hospital patients. In M. Greenblatt, D. Levinson, & G. Klerman (Eds.), *Mental patients in transition: Studies in hospital-community rehabilitation*. Springfield, Ill.: Charles C Thomas, 1961.
- RIKER, B. L. A comparison of methods used in attitude research. *J. abnorm. soc. Psychol.*, 1944, 39, 24–42.
- TAYLOR, J. B. The psychiatrist and the general practitioner. *Arch. gen. Psychiat.*, 1961, 5, 1–6.

(Received January 23, 1963)

ACCURACY AND VARIABILITY OF INFORMATION SOURCES AS DETERMINERS OF PERFORMANCE AND SOURCE PREFERENCE OF DECISION MAKERS¹

JAMES C. NAYLOR

Ohio State University

9 different information sources were examined to determine the relative effects of source accuracy, source variability, and objective expected value upon the source preferences of decision makers. The 9 sources of information were created by using all combinations of 3 levels of mean accuracy of information and 3 levels of error variability. Preference for an information source was shown to be related to both the accuracy of a measure and its variability, with the former appearing to be the most rapidly recognized of the 2. In addition, the objective expected value of a source (as determined by the payoff system used) accounted for nearly all of the variance in group performance.

Executives, either in business or the military, are becoming increasingly dependent upon secondary, rather than primary information in the conduct of their activities. Kidd and Boyes (1959), in defining the role of the executive as one of (a) receiving information from subordinates, (b) integrating this information, and (c) making critical decisions predicated upon the information received, point out several correlates implicit in such a role. Thus, given that a rational decision is one based upon information of some form, many times a decision maker (DM) is not in a position to make direct observations of relevant events himself—instead this function must be delegated to others. These subordinates assume the role of “filterers,” attempting to provide the DM with the most accurate data possible. Indeed, Kidd and Boyes showed that both decision-making speed and accuracy were directly influenced by the amount of distortion (inaccuracy) present in the information supplied by subordinates.

Because of the extremely high costs involved in major industrial and military decisions, emphasis upon redundant information has increased. Thus, it is not uncommon to find

several subordinates (information sources) providing, independently, a DM with the same information. While such overlap may facilitate decisions, given agreement between sources, it can also create ambiguity under conditions of disagreement. Introduction of competing sources of information into the DM's environment makes it necessary for him to learn to discriminate between sources of information in terms of their “goodness” so that when disagreements occur he can select that source most likely to provide accurate information. Since the goodness of an information source is related to the frequency with which the source provides correct data, the magnitude of his error, and the relative costs associated with correct and incorrect decisions based upon his information, the DM must learn to discriminate (a) which source is most apt to be correct, (b) which is more apt to make gross errors, (c) weigh these in terms of costs, and then (d) select the best source as his guide in making a decision.

Given a subordinate who provides a sequence of predictions, the predictions emanating from that person may be considered to represent a distribution (population) of stimuli centering around the point of zero error. Assuming normality, the mean of this distribution is the point of zero error and is expressed by the percentage of correct predictions made by that individual. In addition, the variability of that source will be related to the number and magnitude (if degree of error is important) of the

¹ This research was carried out in the Laboratory of Aviation Psychology and was supported by the Air Force Systems Division, United States Air Force, under Contract No. AF 33(616)-7122, monitored by the 6570th Aerospace Medical Research Laboratories. Permission is granted for reproduction, translation, publication, use, and disposal in whole or in part for any purpose of the United States Government.

incorrect predictions making up the distribution. The problem for the DM thus becomes one of discriminating between distributions which differ in the above manner. This has been called a process of "intuitive statistics" (Edwards, 1961) or "expanded judgment" (Irwin, Smith, & Mayfield, 1956).

The work of Irwin and his collaborators (Irwin & Smith, 1956, 1957; Irwin, Smith, & Mayfield, 1956) has shown that people are able to discriminate quite well between such population means on an intuitive basis and are also able to respond to variability differences. However, their work did not provide any direct comparison of these two abilities, nor did it examine the extent to which the cost function (which is directly related to the mean and variance of a distribution in terms of expected value) of differing distributions plays a role in the DM's choice. The present study was designed to simultaneously explore the effects of mean accuracy, variability, and objective expected value of information sources upon the choice preferences of DMs in a sequential decision task.

METHOD

Subjects. The subjects were 450 female undergraduate students who were enrolled in an introductory psychology course. All were semivolunteers (they had selected this experiment as one to use in fulfilling a course requirement) and all received course credit for participation.

Stimuli. Nine decks of 100 stimuli each were constructed to provide the different experimental conditions. Each stimulus sheet consisted of a 7-inch diameter compass face reproduced on 8.5×11 inch mimeograph paper. Eight different directions were designated on the compass perimeter. These were (in clockwise order) N, NE, E, SE, S, SW, W, and NW. The N point was at the top of the compass face (0°) and all points were separated by 45° . Each sheet provided two predictions of directions—one from a "red" source (drawn in red crayon) and one from a "black" source (drawn in black crayon)—in the form of arrows originating from compass center pointing to one of the eight compass points. Immediately following each stimulus sheet was a knowledge-of-results sheet (KR) that had both the two predictions and the "correct" direction (drawn in blue crayon). Thus, each stimulus deck was comprised of 200 sheets with stimuli and KR sheets being placed in alternate order.

Experimental Design. Nine different experimental conditions were investigated with each condition being defined by the mean accuracy rate and the error variance inherent in the two sources of information being compared. Mean accuracy was defined as the percentage of times (in 100 predictions) an information source was

correct. Three levels, 40%, 50% and 60%, were studied. The variability of a particular information source was defined in terms of the variance (σ^2) of the errors made by a source. Three σ^2 magnitudes were used (small, medium, large) with each of the three mean accuracies, creating a total of nine distributions. Since the compass represents a circular continuum, the greatest error a source could make would result in being off by four categories (e.g., predicting S when N was indeed the correct answer). Errors of less than four units were arbitrarily defined as plus errors if they were in a clockwise direction and as minus if they were counterclockwise. Errors of four units were divided equally as + and - in nature.

Fifty subjects were assigned randomly to each of nine experimental groups. The design used was essentially a constant-stimulus method, where a standard source (Distribution 1), having a mean accuracy of 50% and σ^2 of 1.88, was paired with each of the other eight distributions and with itself. A complete description of the conditions defining the groups is shown in Table 1.

In addition to mean accuracy and variability, it was possible to compute the expected value (EV) and the relative expected value (REV) for each distribution. Two points were given for each correct prediction and points were subtracted for errors in direct relationship to the size of the error (i.e., an error of one unit meant a loss of 1 point, an error of two units meant a loss of 2 points, etc.). Since the distributions were specified, a simple average of points obtained across the 100 predictions provided EV. The REV was defined as: EV comparison distribution - EV standard distribution or the extent to which the comparison source provided a greater payoff than the standard source.

Procedure. The subjects were tested in groups ranging in size from 8 to 20. Each session lasted 1 hour, and only one session was necessary for any subject. All subjects tested at any one time were assigned to the same group and thus received the same experimental deck of predictions. They were instructed that they were participating in a decision-making study, and that they would receive a series of predictions from two sources—a red source and a black source. They were to select each time that source most likely to be correct, and they would receive feedback information about their choice immediately following each selection. They were also told that information sources differed in terms of both accuracy and variability, and these concepts were defined. The scoring procedure was explained and they were requested to try to maximize their total points over the 100 trials.

Each subject was given a mimeographed answer sheet on which to record his decisions (responses). The sheet listed the 100 stimulus events numerically as rows. Three responses were required of a subject on every stimulus event, and provisions were made for these by having four columns, labeled Red, Black, Points, and Sum, immediately to the right of each of the stimulus event row numbers. Each subject recorded his choice (decision) by making a check in the appropriate column (red or black). After the KR sheet was shown, he then scored his choice by putting down the number of points he obtained for that decision under the points column.

TABLE 1
CONDITIONS DESIGNATING EACH OF THE NINE EXPERIMENTAL GROUPS

Group	Standard distribution				Comparison distribution				
	Distribution	Mean accuracy	σ^2	EV	Distribution	Mean accuracy	σ^2	EV	REV ^a
1	1	50	1.88	.16	1	50	1.88	.16	.00
2	1	50	1.88	.16	2	50	.68	.44	.28
3	1	50	1.88	.16	3	50	3.02	-.10	-.26
4	1	50	1.88	.16	4	50	1.90	.42	.26
5	1	50	1.88	.16	5	60	.70	.70	.54
6	1	50	1.88	.16	6	60	3.00	.20	.02
7	1	50	1.88	.16	7	40	1.90	-.14	-.30
8	1	50	1.88	.16	8	40	.66	.18	.02
9	1	50	1.88	.16	9	40	3.00	-.40	-.56

^a Relative expected value.

In addition, he kept a running cumulative total under the sum column.

Two experimenters were present at every testing session. One experimenter stood in the front of the room and presented the stimulus sheets and KR sheets in their prearranged (random) order. The rate of presentation was group paced in that no KR sheet was shown until all subjects had made a response and no stimulus sheet was shown until all subjects had scored their decisions. The other experimenter served as monitor to assure noncollaboration and appropriate timing of presentation. The order of presentation of the sheets was random (Edwards, 1962, found no marked difference between constrained and random orders), but identical for all subjects in a group.

RESULTS

Two dependent variables were measured to assess group performance differences. First, the 100 trials were divided into 10 blocks of 10 trials each and the number of times the subject selected the standard information source was computed for each block, giving a measure of the rate of discrimination between sources. Second, each subject received a total point score based upon his choices over the entire series of 100 trials.

Source Preference. Figure 1 provides a summary of the effect of mean accuracy differences upon the discrimination functions between input sources. Groups 1-3 all had a mean accuracy of 50%, Groups 4-6 all had 60%, and Groups 7-9 had 40%. Apparently, subjects were able to discriminate differences in mean accuracy quite readily, as the slopes representing Groups 4-6 and 7-9 show rapid changes in source preference in the expected directions.

Examination of Figure 2, however, shows that when discrimination of variance differences was involved, subjects had more difficulty in discriminating between sources. Groups 8, 2, and 5 all had less variability than the standard, while Groups 9, 3, and 6 had greater variability. The subjects were clearly able to differentiate

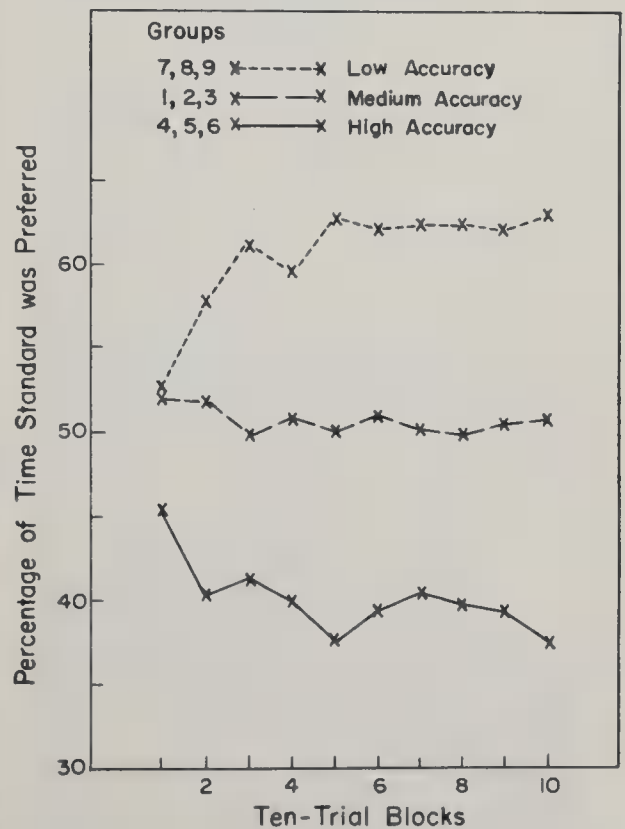


FIG. 1. Group acquisition as a function of differences in the mean accuracy of the comparison input source.

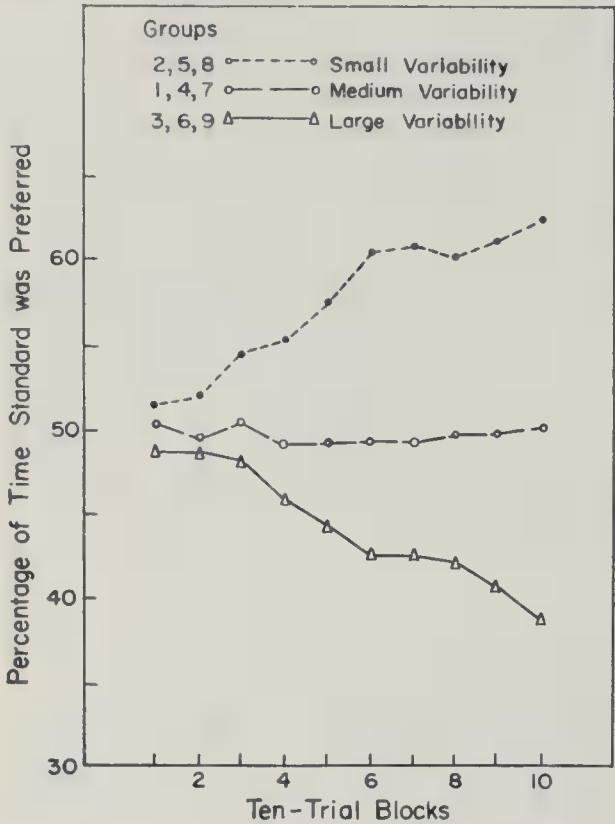


FIG. 2. Group acquisition as a function of differences in the variability of the comparison input source.

between sources according to this parameter, but the slopes are almost linear, indicating a gradual acquisition. This is in sharp contrast to rate of acquisition of mean accuracy differences.

A further illustration of this was given by the individual performances of Group 6 and Group 8. Group 6 had a mean accuracy greater than the standard (60%) but was more variable ($\sigma^2 = 3.00$). These balanced each other in terms of cost, resulting in an REV that was only .04. The performance of Group 6 indicated that during the first five to six blocks subjects were primarily responding to differences in mean accuracy (i.e., the comparison source was clearly preferred), and only during the remaining trials did performance seem influenced by variability (subjects began choosing the standard more often). Much the same finding was found with Group 8. Again, for the first five blocks they appeared to be responding primarily to mean accuracies (by choosing the standard) and during the later trials began to respond to variability differences by selecting the comparison source. It was also interesting to note that final performance of all groups appeared

to be closely related to the REV of that source.

An analysis of variance was performed on the above data and is summarized in Table 2. The analysis provides support for much of the discussion. Both mean accuracy ($F = 66.95, df = 2/441$) and variability ($F = 29.44, df = 2/441$) demonstrated a significant influence on source preferences, and there was no interaction between these two parameters.

It was not unexpected to find the blocks effect nonsignificant, since comparison sources with higher EVs led to a decrease in standard preference while those with lower EVs led to an increase in preference for the standard. This created significant interactions between mean accuracy and blocks ($F = 3.87, df = 18/969$) and between variability and blocks ($F = 7.02, df = 18/969$) which canceled the general blocks effect.

Total Points. The mean number of total points accumulated per subject across the entire 100 trials for each of the nine groups was also examined for significance. The analysis of variance for these data showed that again there was a significant effect on performance due to both mean accuracy ($F = 112.75, df = 2/441$) and variability ($F = 110.98, df = 2/441$) but no significant interaction ($F = 1.13, df = 2/441$). An examination of treatment differences among

TABLE 2
ANALYSIS OF VARIANCE OF SOURCE PREFERENCES (number of times standard source was selected over comparison source in a block of 10 trials) FOR EACH OF THE NINE GROUPS ACROSS THE 10 TRAINING BLOCKS AS A FUNCTION OF DIFFERENCES IN MEAN ACCURACY AND VARIABILITY OF THE TWO INFORMATION SOURCES

Source	df	MS	F
Mean accuracy (A)	2	1557.219	66.95*
Variability (V)	2	684.655	29.44*
A \times V	4	20.889	—
Subjects within groups (Ss/G)	441	23.258	—
Blocks (B)	9	.430	—
B \times A	18	11.656	3.87*
B \times V	18	21.136	7.02*
B \times A \times V	36	1.682	—
B \times Ss/G	969	3.010	—

* $p < .01$.

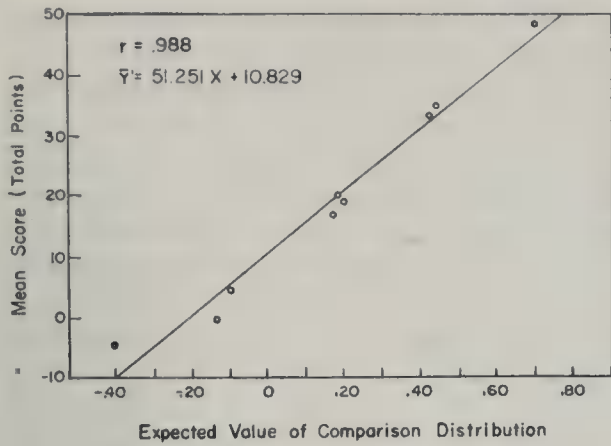


FIG. 3. Relationship between total points accumulated and the expected value of the comparison source.

individual means indicated that observed performance differences were in the expected directions, that is, greater mean accuracy in the comparison source led to better performance, as did less variability.

Since mean accuracy and variability of an information source are the major parameters of the EV of that source, it seemed pertinent to examine to what extent the performance changes brought about by manipulation of these two variables could be explained by EV. A linear regression was computed between total points accumulated by a group and the EV of the comparison source for that group (see Figure 3).² The resulting regression equation $\bar{Y}' = 51.251X + 10.829$, where \bar{Y}' = mean total points and X = EV, resulted in $r = .988$ and $r^2 = .976$, indicating that EV accounted for all

² Since the only difference between EV and REV is the subtraction of a constant (the EV of the standard distribution), the regression equation can also be considered an expression of the relationship between REV and performance.

TABLE 3
ANALYSIS OF REGRESSION FOR TEST OF
GOODNESS OF LINEAR FIT

Source	df	MS	F
Group means	8	15,311	56.49*
Linear	1	119,612	441.37*
Deviation from linear	7	441	1.51
Residual	441	271	
Total	449		

* $p < .01$.

TABLE 4
CORRELATIONS BETWEEN TOTAL NUMBER
OF TIMES STANDARD WAS CHOSEN
AND TOTAL GAME POINTS

Group	r
1	.148
2	-.013
3	.117
4	-.090
5	-.152
6	-.067
7	-.124
8	-.263
9	.280

but .023% of the variance in the subject's performance. It should be noted that each point in Figure 3 is based on 5000 decisions—thus a great deal of stability was anticipated.

As a further test of the linear relationship between EV and performance, an analysis of regression for test of goodness of linear fit was performed as shown in Table 3. As can be seen, the significant overall difference in group means shows a highly significant linear effect and the deviations from this linear hypothesis are not significant, indicating that the hypothesis of a basic linear relationship between performance and EV seems justified.³

Another approach to testing for the adequacy of fit to a particular hypothesis is that given by Grant (1962). Grant's test also resulted in a significant F (F correspondence = 1108.12, $df = 2/7$, $p < .01$) which substantiated the above conclusion that the basic relationship between EV and performance was linear.

Product-moment correlations were computed for subjects within each group between mean total score and the number of times the standard source was preferred (Table 4). In general, the correlations were low but in the expected directions. Those groups having comparison

³ The normal critical values for $F_{lin.}$ and $F_{dev. lin.}$ (using .01 significance level) are $F_{1-\alpha(1/441)}$ and $F_{1-\alpha(7/441)}$. However, since the hypothesis being tested (that of linearity) was obtained from the data being used to test the hypothesis, the test is of an a posteriori nature and a more appropriate critical value is that given by Scheffé (1953). These values are approximated by multiplying the above values by $k - 1$, where k is the number of means (9). Using this correction did not affect the outcome.

sources with lower EVs than the standard source demonstrated positive relationships; that is, the more often they chose the standard, the higher their score. Those groups having higher EVs in the comparison than the standard showed negative correlations; that is, the more often the standard was preferred, the lower their score. Only Groups 1 and 7 failed to show r 's in the expected direction.

DISCUSSION

The finding that both mean accuracy and variability of the information sources affected preference for a source and total score performance was not unexpected. As mentioned earlier, Irwin and his coworkers (Irwin & Smith, 1956, 1957; Irwin, Smith, & Mayfield, 1956) had provided an indication that this was likely. Of particular interest, however, were the differing rates of discrimination by the DMs of these two variables. Apparently, variance of a source is a more difficult concept to acquire than that of means, as the slopes due to variance changes were much more gradual. It should be kept in mind that any attempt to compare rate of discrimination between mean accuracies with rate of discrimination between variances will depend upon the relative magnitudes of the differences to be discriminated. However, the average REV values for those groups having differing variances were $-.27$ for the large variance groups (2, 6, and 9) and $.28$ for the small variance groups (2, 5, and 8), while the REV for the low mean accuracy groups (7, 8, and 9) was $-.27$ and for the high mean accuracy groups (4, 5, and 6) was $.28$. Thus, the relative magnitudes in terms of REV were identical. Also, those groups having both means and variances in the comparison source differing from those of the standard appear to respond initially to the different means and not until later to the different variances, again indicating that in spite of the fact the opposing effects on EV were identical that the initial discrimination was in terms of means rather than variances.

The implications of these data would seem to suggest that variability may be a more critical parameter among information sources than mean accuracy. If, for example, an executive is asked to discriminate among predictions made

by several subordinates, he may have more difficulty doing so if they differ in terms of variability than in terms of likelihood of being correct. In the first case, with both subordinates correct equally often, the tendency of one subordinate to make extreme errors when he is incorrect may be overlooked for some time; that is, a large sample of his behavior may be necessary before the DM begins to realize he tends to be "out in left field" fairly often. This could result in a decrease in decision-making efficiency for some period of time. In the second instance, with both subordinates equally variable but differing in mean accuracy, the DM may pick up this difference very quickly and thus lose little in decision making efficiency. It should certainly be noted that these results relate only to the "either-or" decision situation. The subjects were not allowed to strike a middle ground between predictions given by the several sources. Such an extension is suggested as being an interesting possibility.

The concept of variance as a measure of variability is probably not an intuitively meaningful one to most individuals, at least not to the degree that a mean is an expression of central tendency. It may be that individuals respond to more basic cues, such as extreme scores (range), when asked to judge variability. Since the range is related to sample size (number of trials) because extreme scores have low probabilities of occurrence, initially the range tends to be small, increasing as n increases. Thus, the likelihood of one's perceiving variability differences will increase with the number of trials if he is using the range as an indicant.

The high agreement between EV and performance implies several things. First, apparently mean accuracy and variability, while demonstrated as important variables, are only critical since they in turn affect EV, as performance could be almost exclusively explained in terms of this latter measure. Second, at least for the ranges studied here, DM performance appears to be clearly a linear function of EV regardless of whether EV was manipulated by means or variance. This in turn would seem to imply that EV, as measured by the objective payoffs, was probably a good measure of the subjective utility of the payoff system of the subjects. This is probably more often not the case (Edwards, 1961), and may have occurred

here due to the artificiality of the payoff variable (points). It is doubtful that objective values would be so closely related to subjective values for a business executive.

One final point should be mentioned. Higher relationships between choice preference and total score were expected than were actually obtained. While the correlations were generally in the expected direction, they were quite low, and clearly a large portion of total score variance within a group could not be explained by gross preference frequencies between sources. The distributions were constructed randomly without restrictions (i.e., there was no direct control for "runs" within a source). Thus, there may have been cues of this nature that allowed subjects to deviate from a preferred source with higher EV to a less preferred source with lower EV for periods of short duration and still improve their overall score.

REFERENCES

- EDWARDS, W. Behavioral decision theory. *Annu. Rev. Psychol.*, 1961, **12**, 473-498.
- EDWARDS, W. Dynamic decision theory and probabilistic information processing. *Hum. Factors*, 1962, **4**, 59-73.
- GRANT, D. A. Testing the null hypothesis and the strategy and tactics of investigating theoretical models. *Psychol. Rev.*, 1962, **69**, 54-61.
- IRWIN, F. W., & SMITH, W. A. S. Further tests of theories of decision in an "expanded judgment" situation. *J. exp. Psychol.*, 1956, **52**, 345-348.
- IRWIN, F. W., & SMITH, W. A. S. Value, cost, and information as determiners of decision. *J. exp. Psychol.*, 1957, **54**, 229-232.
- IRWIN, F. W., SMITH, W. A. S., & MAYFIELD, J. F. Test of two theories of decision in an "expanded judgment" situation. *J. exp. Psychol.*, 1956, **51**, 261-268.
- KIDD, J. S., & BOYES, F. Input distortion and observer overlap in decision making. *Mgmt. Sci.*, 1959, **6**, 123-131.
- SCHEFFÉ, H. A. A method for judging all possible contrasts in the analysis of variance. *Biometrika*, 1953, **40**, 87-104.

(Received January 28, 1963)

LIFE GOALS AND VOCATIONAL CHOICE¹

ALEXANDER W. ASTIN AND ROBERT C. NICHOLS

National Merit Scholarship Corporation, Evanston, Illinois

Factor analyses of 56 life-goal and self-rating items were performed on separate samples of 250 male and 250 female college seniors of high ability. 7 identifiable factors were replicated in the 2 analyses: Self-Esteem and Scholarship (primarily associated with self-ratings); Personal Comfort, Prestige, and Altruism (primarily associated with life goals); and Artistic Motivation and Science-Technology (associated both with self-ratings and with life goals). Analyses of 4 life-goal factor scores using 5495 high aptitude students divided into 36 career field groups revealed great differences in the life goals of students pursuing different careers. Results with an open-ended question about life goals supported this finding.

Studies of vocational choice have repeatedly shown that people in one vocational group differ greatly from people in another in their interests, attitudes, and personalities (e.g., Astin, 1958; Clark, 1961; Hammond, 1956; Holland, 1962; Strong, 1943). One area which has not been thoroughly explored is that of differences in life goals and desired accomplishments. One would expect that there are great differences in the goals of people in different occupational groups, since a person's vocation is the means by which he customarily attempts to implement many of his goals. In fact, it is possible to think of vocational choice as a person's attempt to find that work situation which will maximize his chances of achieving the goals which are important to him. Moreover, a knowledge of the goals which are most relevant for various occupations would be of practical value in designing appropriate criteria of success for people in different vocations.

The purposes of this study were to define some of the major life goals of a sample of high aptitude college seniors, and to determine the relationships between the students' goals and their career choices.

METHOD

There were three stages in the study. First, a group of items was constructed which described life goals in three general areas: vocational, personal,

and social. These items were administered to a large sample of high aptitude students at the time of their graduation from college. Next, the students' responses to the items were factored to isolate the general dimensions of life goals and to reduce redundancy in the items. To facilitate the interpretation of these life-goal dimensions, a list of self-ratings was also included in the factor analyses. Finally, groups of students were compared to determine what life goals or areas of achievement are most relevant for the student planning to pursue a particular career.

Items

Twenty-six items pertaining to the student's goals and aspirations were used. They included:

1. "Look ahead 10 years and describe the single future accomplishment which for you would represent achievement or success." (Open-ended response)
2. Highest degree sought
3. Expected income after 10 years
4. Hoped-for vocational level: "Compared with other people in my chosen vocation, I *hope* to be (check one): average; above-average; in top 25%; in top 10%; in top 5%, in top 1%."
- 5-26. Specific life goals in three broad areas:

Vocational— inventing an apparatus or piece of equipment; making a contribution to scientific knowledge; becoming an authority on a special subject in my field; having a musical composition played or published; becoming an expert in finance and commerce; having poems, novels, or short stories published; becoming an outstanding instrumental musician or singer; producing original painting, sculpture, etc.

Social—helping others who are in difficulty; being a good parent; becoming a community leader; becoming influential in public affairs; contributing to human welfare.

Personal—being well liked; financial security; becoming happy and content; having the time and means to relax and enjoy life; finding a real

¹ This study is a part of the research program of the National Merit Scholarship Corporation and was supported by grants from the National Science Foundation, the Carnegie Corporation, and the Ford Foundation.

purpose in life; obtaining awards and recognition; becoming famous; becoming a mature person; following a formal religious code.

Each of these 22 specific life goal items was rated by the subject on a 4-point scale:

- Essential* (something I must achieve) (score 4)
- Very Important* to achieve (but not essential) (score 3)
- Somewhat Important* for me to achieve (score 2)
- Of Little or No Importance* for me (score 1)

The subject was also asked to rate himself on the following 29 traits: absentmindedness, aggressiveness, artistic ability, athletic ability, cheerfulness, clerical ability, conservatism, drive to achieve, emotional stability, expressiveness, independence, leadership ability, mathematical ability, mechanical ability, neatness, originality, perseverance, popularity, popularity with opposite sex, practical-mindedness, reading ability, scholarship, self-confidence, self-understanding, sociability, speaking ability, stubbornness, understanding of others, and writing ability. Students checked trait on a 4-point scale:

- Top ten per cent (score 4)
- Above average (score 3)
- Average (score 2)
- Below average (score 1)

On the basis of a content analysis of 200 questionnaires, two codes for the open-ended question were developed. The first code indicated whether the general content of the subject's response was Vocational (e.g., "becoming a good physician," "publishing a book"), Personal (e.g., "becoming more mature," "knowing myself better"), or in the category of Family or Marriage (e.g., "raising my children intelligently," "being a good provider"). (A given response could be classified in more than one of these general categories.) The second code indicated the specific nature of some of the vocational goals: that is, whether they were Technological ("building a bridge"), Intellectual ("making an important scientific discovery"), Artistic ("writing good poetry"), Social ("benefiting humanity in some way"), or concerned with Status or Power ("making a lot of money"). Many of the Vocational responses—for example, "becoming a competent teacher," "doing a good job"—could not be classified in one of these five categories, because they suggested no *specific* vocational achievement. In another sample of 300 questionnaires, agreement between two independent judges in classifying responses to the open-ended question was 84%.

Subjects

Questionnaires which included the 55 items were mailed to 8489 students—Merit Finalists and recipients of the Letter of Commendation from the 1957 National Merit Scholarship competition—in the Spring of 1961, shortly before they graduated from college. A 78% rate of return yielded 6593

questionnaires, a total further reduced to 5495 when questionnaires which were incomplete or which contained unclassifiable responses (primarily rare career choices) were eliminated. The sample of 5495 included 3830 men and 1665 women. Although the exact biases in the sampling are unknown, these subjects can perhaps be best characterized as a national sample of students of very high academic aptitude (for men and women, respectively, mean SAT Verbal scores were 651 and 669, and mean SAT Mathematical scores were 676 and 629).²

Factor Analyses

Two-hundred and fifty male and 250 female subjects were drawn at random from the larger sample. Product-moment correlations among the 54 life-goal and self-rating items were computed separately for the two sexes. To assess the importance of responses biases, two other variables were included in the matrix: the total number of specific life goals checked "essential," and the total number of self-rating items checked "in the top ten percent." The two resulting 56×56 matrices were factored separately, using the principal components method with one iteration for convergence of communalities. Factors accounting for more than 3% of the variance were retained and rotated to the normalized varimax criterion of orthogonal simple structure. Computations were performed on an IBM 704 computer using a program written by Wexler (1959).

Life Goals in Different Career Areas

The 3830 male subjects were divided into 19 different groups on the basis of their expressed career choices. The 1665 female subjects were similarly divided into 17 groups. The *N*s in each of the 36 groups ranged from 20 to 615. (Career choice groups which had fewer than 20 subjects had been excluded earlier.) There were actually only 24 different careers represented by the 36 groups, since men and women had 12 careers in common. The factor scores derived from the specific life-goal items and the coded responses to the open-ended question of each of the 36 groups were then compared.

RESULTS

Eight factors were obtained from the analyses of both the male and female matrices.³ The first seven of these factors, which

² A more detailed description of the sample has been given elsewhere (Astin, 1963).

³ A 14-page list containing the correlation matrices and rotated factor matrices for men and women has been deposited with the American Documentation Institute. Order document No. 7752 from ADI Auxiliary Publications Project, Photoduplication Service, Library of Congress, Washington, D. C. 20540. Remit in advance \$1.75 for micro-

TABLE 1
FACTORS OBTAINED FROM ANALYSES OF MALE AND FEMALE MATRICES

Factor	Loading		Factor	Loading	
	Male analysis	Female analysis		Male analysis	Female analysis
I. Self-Esteem			Reading ability	.17	.37
Life Goals			Understanding of others	.20	.35
Becoming a community leader	.49	.39	Percent variance accounted for	4.5	5.7
Becoming influential in public affairs	.49	.45	IV. Scholarship		
Becoming an expert in finance and commerce	.44	.23	Life Goals		
Expected income after 10 years	.33	.11	Hoped-for vocational level	.39	.38
Self-Ratings			Becoming an authority on a special subject in my field	.38	.20
Popularity	.66	.68	Self-Ratings		
Leadership ability	.64	.46	Drive to achieve	.62	.69
Popularity with opposite sex	.64	.57	Perseverance	.55	.23
Sociability	.63	.77	(Number of self-ratings in top 10%)	(.55)	(.61)
(Number of self-ratings in top 10%)	(.61)	(.47)	Independence	.45	.37
Self-confidence	.60	.40	Scholarship	.44	.60
Emotional stability	.54	.47	Originality	.39	.30
Cheerfulness	.54	.58	Self-confidence	.37	.50
Aggressiveness	.51	.27	Leadership ability	.27	.50
Expressiveness	.44	.39	Speaking ability	.18	.39
Independence	.43	.15	Aggressiveness	.25	.35
Understanding of others	.42	.36	Percent variance accounted for	4.8	6.3
Practical-mindedness	.42	.12	V. Science-Technology		
Athletic ability	.39	.31	Life Goals		
Speaking ability	.39	.39	Making a contribution to scientific knowledge	.64	.64
Self-understanding	.36	.38	Inventing an apparatus or piece of equipment	.56	.43
Percent variance accounted for	10.3	7.3	Becoming an authority on a special subject in my field	.18	.37
II. Personal Comfort			Self-Ratings		
Life Goals			Mechanical ability	.65	.55
Having the time and means to relax and enjoy life	.70	.64	Mathematical ability	.50	.48
Becoming happy and content	.67	.66	Percent variance accounted for	3.7	3.5
(Total number of goals rated essential)	(.67)	(.49)	VI. Prestige		
Financial security	.66	.61	Life Goals		
Being well-liked	.55	.55	Becoming famous	.71	.62
Becoming a mature person	.53	.16	Obtaining awards and recognition	.64	.56
Being a good parent	.50	.19	Becoming influential in public affairs	.52	.22
Becoming an expert in finance and commerce	.31	.13	Becoming a community leader	.42	.01
Highest degree sought	-.23	-.30	Becoming an expert in finance and commerce	.36	.05
Self-Rating			Becoming an authority on a special subject in my field	.35	.45
Conservatism	.32	.23	Highest degree sought	.13	.37
Percent variance accounted for	6.4	4.3	Self-Rating		
III. Artistic Motivation			Practical-mindedness	.15	.36
Life Goals			Percent variance accounted for	4.2	3.6
Having poems, novels, or short stories published	.63	.63	VII. Altruism		
Having a musical composition played or published	.47	.18	Life Goals		
Producing original painting, sculpture, etc.	.36	.50	Contributing to human welfare	.78	.71
Becoming an outstanding instrumental musician or singer	.36	.20	Helping others who are in difficulty	.66	.65
Self-Ratings			Finding a real purpose in life	.47	.54
Artistic ability	.52	.66	(Total number of goals rated essential)	(.46)	(.74)
Writing ability	.48	.50	Becoming a mature person	.44	.42
Originality	.46	.70	Becoming a community leader	.30	.43
Expressiveness	.36	.51	Following a formal religious code	.28	.36
(Number of self-ratings in top 10%)	(.24)	(.45)	Being a good parent	.30	.32
			Percent variance accounted for	4.5	5.1

seem to be common to the two analyses, are discussed below, and variables with loadings of .30 or larger in either analysis are listed (see Table 1).

Factor I appears to be primarily a self-concept factor; the self-ratings reflect a general tendency on the part of the subject to perceive himself favorably. In particular, he emphasizes his interpersonal and social competence. The life goals which load on this factor involve attaining power and positions of leadership.

Factor II, Personal Comfort, is concerned almost exclusively with life goals. High-loading items on this factor appear to represent the "good life" syndrome: being comfortable, happy, well liked, and financially secure. Conservatism is the only self-rating which loads on this factor.

Factor III is clearly concerned with artistic endeavors. It combines life goals in the area of artistic achievement with high self-ratings of artistic and creative talent. For the women, however, two life goals associated with musical achievement (composing and performing) had only negligible loadings on this factor, and instead formed a separate factor. This separate music factor had no counterpart in the male analysis, and so is not reported in detail.

Factor IV, another self-rating factor, contains several self-ratings which in previous studies (Holland & Astin, 1962) were shown to predict academic achievement.

Factor V, Science-Technology, includes both life goals and self-ratings among the high-loading items. Like the Artistic Motivation factor, it combines motivation to achieve in a specific field with perceived ability in that field.

Factor VI, Prestige, is primarily a life-goal factor. High-loading items indicate a striving for social recognition and, among men, for social power (i.e., having "influence" and being a "leader").

Factor VII, Altruism, is exclusively concerned with life goals. High-loading items involve helping others and achieving personal

maturity. These goals suggest a highly developed super ego.

The scores for four life-goal factors (Personal Comfort, Science-Technology, Prestige, and Altruism) were computed by summing the raw scores on the highest-loading life-goal items. The two highest-loading items were used to compute factor scores for Science-Technology, Prestige, and Altruism; the three highest-loading items were used in computing the Personal Comfort score. (The Artistic factor, which also had high loadings from life-goal items, was not used because of the differences in the size of the loadings between the male and female analyses.)

Table 2 shows the correlations among the four life-goal factor scores separately for the two sexes. (The 1091 men and 457 women were drawn from the larger sample in connection with another study; the 500 subjects used in the two factor analyses have been excluded.) With the possible exception of the correlation between the Personal Comfort and Prestige factor scores in the male sample ($r = .21$), the four factor scores appear to be relatively independent.

To compare the 36 career groups in terms of the four life-goal factors, the percentages of subjects in each group whose scores on the life-goal items comprising the factors averaged 3 ("very important") or higher were computed. These percentages are shown in Table 3, together with the median "expected income after 10 years" as reported by the subjects in each group.

All five variables shown in Table 3 discriminate significantly among the 36 groups. Of the 10 $2 \times k$ chi square tests (high scorers

TABLE 2

INTERCORRELATIONS OF FOUR LIFE-GOAL
FACTOR SCORES

Life-Goal Factor	1	2	3	4
1. Science-Technology		.01	.08**	-.07*
2. Altruism	-.06		.08**	.01
3. Prestige	.07	.09*		.21**
4. Personal Comfort	-.06	.01	.05	

Note.—Factor score correlations for males ($N = 1,091$) are shown in the upper-right triangular matrix; correlations for females ($N = 457$) are shown in the lower-left matrix,

* $p < .05$.

** $p < .01$.

TABLE 3
LIFE GOALS OF 36 CAREER CHOICE GROUPS: STRUCTURED ITEMS

Career choice group	N	Percent obtaining high scores ^a on each of the following life-goal factors				Median expected income after 10 years
		Personal comfort	Science-Technology	Prestige	Altruism	
Males						
School teacher (science)	42	73	07	12	85	\$ 7,636
School teacher (nonscience)	73	56	06	14	87	7,697
College professor (science)	266	57	52	14	68	8,693
College professor (nonscience)	397	58	08	28	74	8,037
Writer or journalist	47	50	02	24	49	8,100
Undecided	432	62	21	17	58	9,731
Physician	377	68	31	15	94	13,402
Lawyer	374	82	02	39	78	13,017
Research biologist	39	54	77	13	67	8,453
Research chemist	142	57	79	23	59	11,928
Research mathematician	73	56	64	21	47	12,904
Research physicist	258	63	79	22	54	12,322
Research social scientist	20	68	84	30	70	9,545
Research engineer	102	58	72	17	70	13,524
Engineer	615	83	62	14	56	12,835
Clergyman	103	17	02	09	100	6,163
Military officer	129	63	27	28	74	8,363
Business executive	307	91	10	26	58	14,111
Architect	34	56	09	25	66	12,207
Total	3830	68	35	21	68	11,297
Females						
School teacher (science)	77	76	08	04	91	6,667
School teacher (nonscience)	388	81	01	07	90	7,060
College professor (science)	70	59	18	14	79	7,881
College professor (nonscience)	250	68	04	11	83	7,133
Writer or journalist	42	69	00	05	67	7,501
Undecided	212	72	07	09	71	7,141
Physician	47	56	29	04	93	10,973
Lawyer	20	60	00	11	84	11,138
Research biologist	33	61	45	06	78	8,000
Research chemist	81	73	57	10	76	8,565
Research mathematician	27	69	35	05	69	8,929
Research physicist	20	58	58	15	55	10,251
Homemaker	224	82	02	03	73	7,586
Medical technician	35	83	09	03	82	7,632
Nurse	50	94	06	06	98	4,688
Secretary or clerk	48	88	00	02	73	6,429
Social worker	41	68	02	00	100	7,414
Total	1665	75	09	07	81	7,356

^a A "high score" is defined as one where the subject's mean score on the specific items comprising the factor is 3 (very important) or higher.

versus low scorers across k career groups), all were significant at the .001 level except for the test of the Prestige factor among women ($\chi^2 = 30.02$, $df = 16$, $p < .02$).

The differences between the sexes as to life goals are immediately apparent. Men tend to be more concerned with achievement in science and technology and with gaining pres-

tige. Men also expect to make more money than do women. Women, on the other hand, are more concerned with altruism and with personal comfort. Some of these sex differences are reversed, however, when career choice is taken into account: For example, women planning to be physicians, lawyers, and research physicists are *less* concerned with achieving personal comfort than are males planning the same careers.

Even in absolute terms, the discrepancies in life goals between some of the career groups are very large. Some of these differences are illustrated in Figure 1, which shows distribu-

tions of scores on the five variables for selected groups. It is clear from Figure 1 that there is practically no overlap on some life goals among certain career choice groups.

Of perhaps equal importance is the finding that, *within* some of the extreme groups shown in Figure 1, there is still considerable variability on certain life goals. For example, even though men planning to become lawyers are more highly motivated toward prestige than are any other career choice group, 16% of these students actually obtain the lowest possible score (2) on the prestige factor.

Coded responses to the open-ended life-

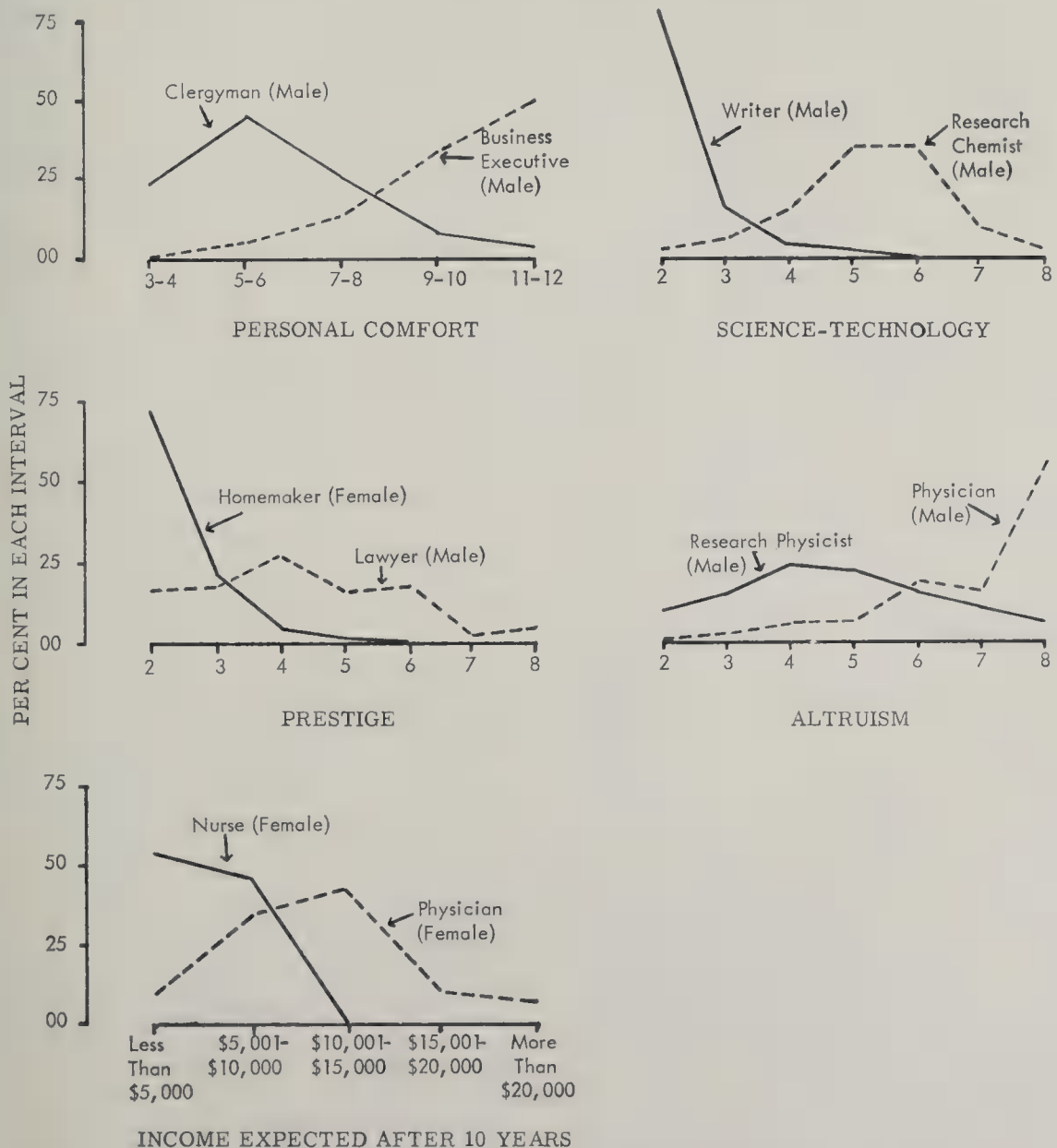


FIG. 1. Distributions of selected career field groups on five life goals.

TABLE 4
LIFE GOALS OF 36 CAREER CHOICE GROUPS: FREE RESPONSE

Career choice group	N	Percent giving responses in each of the following general categories			Percent reporting each of the following specific types of vocational goals				Status and power
		Voca- tional	Family or mar- riage	Personal	Techn- ical	Intel- lectual	Artistic	Social	
Males									
School teacher (science)	42	83	02	10	00	07	00	24	12
School teacher (nonscience)	73	88	15	11	00	01	03	18	20
College professor (science)	266	82	10	11	00	23	01	07	21
College professor (nonscience)	397	86	10	13	00	17	15	10	21
Writer or journalist	47	79	04	11	00	06	32	06	32
Undecided	432	69	15	25	01	10	07	07	18
Physician	377	87	13	15	01	10	00	14	14
Lawyer	374	81	18	16	00	01	00	06	49
Research biologist	39	80	13	13	03	48	00	05	05
Research chemist	142	81	16	10	04	36	01	05	16
Research mathematician	73	81	11	15	07	37	04	00	11
Research physicist	258	83	11	12	02	44	00	03	13
Research social scientist	20	90	10	00	10	65	05	05	05
Research engineer	102	83	15	13	10	27	01	01	16
Engineer	615	81	14	16	06	08	00	03	37
Clergyman	103	79	03	18	00	00	02	44	06
Military officer	129	87	08	10	01	02	02	02	60
Business executive	307	79	26	21	00	00	01	04	54
Architect	34	79	06	21	18	03	06	03	35
Total	3830	81	14	16	02	14	03	08	28
Females									
School teacher (science)	77	62	58	16	00	01	01	22	04
School teacher (nonscience)	388	63	52	15	00	01	04	18	04
College professor (science)	70	80	36	09	00	09	00	15	06
College professor (nonscience)	250	74	41	16	00	08	09	13	05
Writer or journalist	42	74	69	17	00	00	29	00	07
Undecided	212	60	41	23	00	05	06	09	04
Physician	47	89	27	17	00	08	00	13	08
Lawyer	20	90	60	10	00	00	05	25	30
Research biologist	33	49	42	12	03	16	03	03	06
Research chemist	81	74	42	17	02	23	00	07	07
Research mathematician	27	74	59	07	00	07	00	00	11
Research physicist	20	68	26	21	05	11	00	11	16
Homemaker	224	18	88	09	00	00	02	03	00
Medical technician	35	49	77	11	00	06	00	03	03
Nurse	50	46	76	06	00	00	02	04	12
Secretary or clerk	48	45	62	19	00	02	02	04	13
Social worker	41	70	58	25	00	00	03	35	03
Total	1665	60	54	15	00	05	04	12	05

Note.—To the item, "Look ahead 10 years and describe the single future accomplishment which for *you* would represent achievement or success."

goal question are shown in Table 4. Men are more likely than women to give a response with vocational content, whereas women are much more likely to give a response with family or marriage content. When career choice is controlled, differences between men and women in the percentages of vocational responses disappear in some instances (e.g., college professors of science and physicians).

However, women consistently exceed men in the percentages of family or marriage responses, regardless of career choice. There appear to be no sex differences with respect to the percentages of subjects giving personal responses to the open-ended question. The vocationally undecided subjects of both sexes give a high proportion of personal responses. The results for specific vocational responses

(last 5 columns of Table 4) were generally consistent with our expectations. Codes for intellectual and for status and power goals appear to differentiate best among the career groups, with the several groups of scientific researchers giving the highest proportion of intellectual responses, and the military officer, business executive, and lawyer groups giving the highest proportions of responses concerned with status and power.

DISCUSSION

It is apparent from these results that high aptitude students in different career fields have very different types of life goals. Furthermore, there are some goals which appear to be entirely atypical for certain career groups.

The great variance in life goals among these students has important implications for the construction of criterion measures of later adult achievement. In prediction studies, particularly, it is necessary that the criterion measures be appropriate. For example, it does not make sense to use income as an indication of "success in adult life," when to many people money—beyond the level necessary for subsistence—is of little importance. Unless criterion measures are relevant to the subjects' actual goals, the skills and experiences necessary for successful performance may well be overlooked, and the best predictors of success might turn out to be merely correlates of the subject's original intentions.

We can illustrate this problem with an extreme example. Suppose we designate, as one index of adult success, the degree of personal comfort which a person is able to achieve. It is very likely that, 10 or 15 years after college, students who initially chose careers as business executives will score much higher on this criterion than those who chose careers as clergymen. Now if this criterion is applied to all students, the ability, personality, and biographical items which differentiate business executives from clergymen will probably turn out to be the best predictors. As a result, less weight will be given to those variables which predict the success of those people *within* the business executive group who were initially motivated to achieve personal com-

fort. An even more serious consequence is that the variables which predict achievement in a given direction for an undifferentiated sample will turn out to have the opposite relationship to achievement within an occupational group striving for the particular criterion.

One unexpected result was the failure of "expected income after 10 years" to load on any of the life-goal factors, even though "financial security" had a high loading on Personal Comfort. This result cannot be attributed to the students' errors in estimating, since there were great—and generally realistic—differences among the median expected incomes of the various groups. Perhaps the amount of money per se which the student expects to make is not a major motivating factor in his choice of a career.

Factors similar to the ones identified in these factor analyses may prove to be useful in developing an empirical scheme for classifying careers. Different careers could be compared on the basis of the similarities and differences in the life goals of people who pursue the careers. For example, the present results suggest that students who are planning careers as military officers, business executives, and lawyers are similar in that they all have a relatively strong desire for prestige (life-goal factor) and for status and power (open-ended question). Similarly, women who choose the career of homemaker, medical technician, nurse, and secretary or clerk are similar in that they express a great desire for personal comfort (life-goal factor) and have a relatively high interest in goals concerned with family or marriage (open-ended question).

This study has demonstrated that there are great differences in the goals toward which people in different career groups are striving. Since the subjects were just beginning their careers, their goals had not yet been influenced greatly by their experiences in their careers, and so it is reasonable to suppose that many of the differences in goals between career groups exist prior to career choice. Although this interpretation needs more confirmation than the present study provides, it still seems likely that life goals are important determiners of career choice and possibly of

satisfaction with the career once a choice has been made.

REFERENCES

- ASTIN, A. W. Dimensions of work satisfaction in the occupational choices of college freshmen. *J. appl. Psychol.*, 1958, 42, 187-190.
- ASTIN, A. W. Differential college effects on the motivation of talented students to obtain the PhD. *J. educ. Psychol.*, 1963, 54, 63-71.
- CLARK, K. E. *Vocational interests of nonprofessional men*. Minneapolis: Univer. Minnesota Press, 1961.
- HAMMOND, MARJORIE. Motives related to vocational choices of college freshmen. *J. counsel. Psychol.*, 1956, 3, 257-261.
- HOLLAND, J. L. Some explorations of a theory of vocational choice: One- and two-year longitudinal studies. *Psychol. Monogr.*, 1962, 76(26, Whole No. 545).
- HOLLAND, J. L., & ASTIN, A. W. The prediction of the academic, artistic, scientific, and social achievement of undergraduates of superior scholastic aptitude. *J. educ. Psychol.*, 1962, 53, 132-143.
- STRONG, E. K. *Vocational interests of men and women*. Stanford: Stanford Univer. Press, 1943.
- WEXLER, J. *Multiple analysis program system*. Tempe: Arizona State University, 1959.

(Received January 28, 1963)

VALUE AND PERSONALITY DIFFERENCES BETWEEN OFFENDERS AND NONOFFENDERS¹

ROBERT R. KNAPP²

United States Navy Medical Neuropsychiatric Research Unit, San Diego, California

The Socialization (So) scale of the California Psychological Inventory (CPI), the DF Opinion Survey, and a measure of 6 interpersonal values were administered to an offender and a nonoffender Navy enlisted sample to investigate any differences in values held by these groups, independent of the usually discriminant variables of verbal aptitude and education. The CPI So scale, 3 scales from the DF Opinion Survey, and 2 scales from the measure of interpersonal values differentiated significantly between the 2 groups. The present Navy offender sample was characterized as having attitudes favorable toward escapism and toward nonconformity to rules and regulations, and as being lower on a continuum of socialization.

That differences in values exist between individuals who fail to conform to the rules and regulations of their immediate environment, as contrasted with those who do conform, would seem to be an obvious hypothesis. The demonstration of such differences through the use of psychometric devices might not follow so readily. The purpose of the present study is to investigate the differences in professed values between those who achieve a relatively satisfactory adjustment to their environment as contrasted with those who fail, or evidence difficulty, in achieving such an adjustment.

The determination of value and other personality correlates of nonconforming, or delinquent, behavior is the first step to be taken toward understanding the potentially habitual or frequent offender. From such an understanding, programs for selection, for placement, or for remedial and rehabilitative efforts can be more meaningfully directed toward their respective goals. The military situation provides a most advantageous environment for the isolation of factors associated with maladjustive behavior. The magnitude of the problem within the Navy has been reviewed by Courtney and Jones (1960) and by Wilkins (1961).

Previous investigations have suggested that value dimensions are related to delinquency. Gordon (1961), for example, has noted that delinquent groups differ consistently and significantly from nondelinquent groups on the Conformity scale of the Survey of Interpersonal Values (SIV; Gordon, 1960). Knapp (1963) also found that Conformity scores were significantly related to a rate of delinquency criterion within a Navy delinquent sample. Further, in the latter study, the Socialization (So) scale of the California Psychological Inventory (CPI; Gough, 1957) was found to be significantly related to the delinquency rate criterion. The development of that scale is founded upon the notion that a continuum of socialization exists along which individuals vary and that the determination of a person's position on this continuum can be ascertained by psychological measurement (Gough, 1960; Gough & Peterson, 1952).

METHOD

Procedure. To test the significance of difference between values held by those having a record of disciplinary offenses as compared with those having no such record, two groups of subjects were selected from aboard a Navy ship. The groups were selected such as to be as nearly equated as was feasible on the variables, education, verbal aptitude, and length of service, which have been previously demonstrated to be associated with delinquency potential. Educational level has been shown to be related to tendency toward disciplinary offenses in the service (Flyer, 1959; Knapp, 1963). Some investigators (e.g., Flyer, 1959) have also found unsuitable military groups

¹The opinions and conclusions expressed are those of the author and do not necessarily reflect the views of the Navy Department.

²Now with Educational and Industrial Testing Service, San Diego, California.

to be lower in verbal aptitude although this variable has not always been found to be related to military delinquency (Knapp, 1963). Also, since a man's length of service is important in terms of opportunity to get into trouble, the groups were approximately equated for length of service. Further, there was no significant difference in mean age between the two groups.

In selecting the two groups, as stated above, an attempt was made to equate them as nearly as possible on relevant variables. But since a perfect matching was not feasible, analysis of covariance was used, primarily to reduce the error variance, thus providing a more sensitive test of differences between groups on the variables under study, as well as, in part, to control for differences remaining between groups on the control variables.

Subjects. The subjects were 82 Navy enlisted men assigned to one of the Navy's heavy cruisers. They ranged in ages from 17 to 25 with a mean of 19.9 years. Educational level ranged from 7 to 11 with a mean of 9.5 years of school completed. General Classification Test (GCT) scores ranged from 30 to 58 with a mean of 44.7. Length of service ranged from 7 to 74 months with a mean of 32.1 months of service. In comparing the present sample with a sample of 11,000 incoming recruits for the year 1960,³ the present mean of 9.5 years

³ A. E. McMichael and J. A. Plag, personal communication, 1962.

of school completed contrasts with an approximate mean of 10.7 for the recruit population and the mean GCT of 44.7 contrasts with an estimated mean of 50.0 for the recruit population. Previous studies (Flyer, 1959; Force & Meyer, 1959) have found lower intelligence and non-high-school graduation to be associated with separation from the service for unsuitability reasons. Thus, the present sample is primarily from this segment of the military population expected to have the highest incidence of offenders.

Subjects were divided into offender and non-offender groups on the basis of examination of each man's service record. Men having any offenses in their record were classified as "offenders" for present analysis. Offenses committed consisted primarily of unauthorized absence. Other offenses were primarily military in nature such as "disrespect" and "failure to obey a lawful order." The groups used for the present analysis consisted of 36 offenders and 46 nonoffenders.

Inventories. The DF Opinion Survey (Guilford, Christensen, & Bond, 1956), a 300-item Yes-No instrument yielding measures of 10 dynamic, or motivational factors, was administered. The value dimensions measured by a second instrument are also considered in the present study. This instrument yielded normative (Yes-No) measures of the dimensions measured by the forced-choice SIV (Gordon, 1960).

TABLE 1
ANALYSIS OF COVARIANCE SUMMARY TABLE FOR COMPARISON OF OFFENDER AND NONOFFENDER GROUPS

Value & personality dimensions	Adjusted means		Mean squares		F
	Offenders	Nonoffenders	Error	Groups	
DF Opinion Survey					
Need for Attention	17.72	14.61	39.67	190.36	4.80*
Liking for Thinking	14.76	13.23	40.01	45.44	1.14
Adventure vs. Security	19.73	16.19	25.43	244.84	9.63**
Self-Reliance vs. Dependence	15.92	14.74	32.04	27.16	.85
Aesthetic Appreciation	11.72	10.03	61.03	56.28	.92
Cultural Conformity	14.84	15.98	24.03	25.39	1.06
Need for Freedom	15.57	12.30	29.75	210.01	7.06**
Realistic Thinking	12.72	14.87	34.10	89.87	2.64
Need for Precision	10.94	10.07	57.93	15.15	.26
Need for Diversion	14.71	12.47	36.52	97.93	2.68
Dimensions reflected from SIV					
Support	7.96	9.25	19.99	32.96	1.65
Conformity	8.85	11.14	21.70	102.68	4.73*
Recognition	6.61	5.70	14.29	16.33	1.14
Independence	12.00	9.57	14.19	114.47	8.07**
Benevolence	5.77	7.36	22.09	49.50	2.24
Leadership	6.50	5.68	19.75	13.13	.67
CPI					
So Scale	25.62	30.49	54.07	465.44	8.61**

* $p < .05$.
** $p < .01$.

TABLE 2
INTERCORRELATIONS AMONG INVENTORY SCALES

Values	Normative (SIV)						DF Opinion Survey										CPI	
	S	C	R	I	B	L	NA	LT	AS	SR	AA	CC	NF	RT	NP	ND	SO	
Normative (SIV)																		
S Support		36	66	-12	47	07	33	21	-05	-29	08	34	-20	-32	17	28	15	
C Conformity			07	-31	59	-08	08	16	-18	-19	12	57	-46	-01	17	17	40	
R Recognition				04	22	35	53	11	15	-15	18	09	04	-40	00	23	-05	
I Independence					-38	16	10	-28	16	-08	-03	-09	44	-27	-15	-06	-32	
B Benevolence						-20	06	21	00	-10	19	37	-44	-07	18	29	28	
L Leadership							24	09	-06	03	00	-04	16	06	07	-13	-07	
DF Opinion Survey																		
NA Need for Attention							30	28	-20	23	33	34	-63	29	47	07		
LT Liking for Thinking								32	10	47	39	-20	-11	74	36	14		
AS Adventure vs. Security									24	22	03	13	-26	32	28	-17		
SR Self-Reliance vs. Dependence										-14	-14	-22	48	02	01	14		
AA Aesthetic Appreciation											13	-11	-17	42	37	-07		
CC Cultural Conformity												-04	-21	45	34	22		
NF Need for Freedom													-49	-11	07	-41		
RT Realistic Thinking															-08	-39	13	
NP Need for Precision																45	15	
ND Need for Diversion																		14
CPI																		
SO Socialization																		

Note.—N = 82.

To obtain scores for these variables not confounded by the forced-choice technique, each item stem in the SIV was prefaced with the phrase "It is important to me." Subjects were instructed to answer Yes or No to each item. An individual's score on each scale was the sum of his Yes responses on that scale. Previous results demonstrated that the normative measure of these value dimensions possessed greater validity than the standard forced-choice, or ipsative, instrument for differentiating between the offender and nonoffender groups. Since the present report is concerned more with determining the relationship of *dimensions* to offense criteria rather than particular instruments per se, only the normative data are presented here.

In addition to these inventories, the So scale of the CPI (Gough, 1957) was administered. This scale was developed specifically to predict delinquency and has been shown to be significantly related to an offense criterion in the Navy (Knapp, 1963).

RESULTS

Table 1 presents the data from the analysis of covariance showing the *F* ratio for each of the value dimensions and Table 2 presents intercorrelations among the scales based on the present sample of 82. From examination of Table 1, it can be seen that 3 of the 10 dimensions measured by the DF Opinion Survey significantly differentiated between the offender and nonoffender groups. Of the 6

dimensions measured by the normative instrument designed to reflect SIV values, 2 significantly differentiated the groups.

The CPI So scale also differentiated between the groups at a high level of significance and in the expected direction.

DISCUSSION

The present results from the DF Opinion Survey depict the Naval offender as having a greater need for attention, adventure, and freedom. The value dimensions measured by the normative instrument designed to reflect SIV value dimensions depict the offender as placing less importance on conformity and greater importance on independence.

The findings are consistent with common sense observations. However, it is noteworthy that personality measures, based on simple self-report techniques can in fact, differentiate between groups equated for the variables so often found to differentiate the delinquent from the nondelinquent; that is, educational level and verbal aptitude.

Results suggest that, among the present lower education, lower GCT group, attitudes favorable to escapism (as reflected in present

measures of the dimensions of freedom, independence, adventure) when combined with nonconformist attitudes toward regulations are related to the tendency towards delinquent behavior within the Navy structure. The present study illustrates that psychometric instruments can be called upon to shed more light on the dynamics of the military offender.

REFERENCES

- COURTNEY, D., & JONES, N. W. JR. Research approach to the Naval offender problem in the U. S. Navy. Report No. 34, 15 May 1960, Courtney and Company, Philadelphia, Pennsylvania, Contract Nonr-2845(00), Office of Naval Research.
- FLYER, E. S. Factors relating to discharge for unsuitability among 1956 airman accessions to the Air Force. *USAF WADC tech. Rep.*, 1959, No. 59-201.
- FORCE, R. C., & MEYER, J. K. Prediction of separation of Air Force trainees. *J. psychol. Stud.*, 1959, 11, 28-31.
- GORDON, L. V. *Survey of Interpersonal Values*. Chicago, Ill.: Science Research Associates, 1960.
- GORDON, L. V. Conformity among the nonconformists. *Psychol. Rep.*, 1961, 8, 383.
- GOUGH, H. G. *California Psychological Inventory*. Palo Alto, Calif.: Consulting Psychologists Press, 1957.
- GOUGH, H. G. Theory and measurement of socialization. *J. consult. Psychol.*, 1960, 24, 23-30.
- GOUGH, H. G., & PETERSON, D. R. The identification and measurement of predispositional factors in crime and delinquency. *J. consult. Psychol.*, 1952, 16, 207-212.
- GUILFORD, J. P., CHRISTENSEN, P. R., & BOND, N. A. *The DF Opinion Survey*. Beverly Hills: Sheridan Supply, 1956.
- KNAPP, R. R. Personality correlates of delinquency rate in a Navy sample. *J. appl. Psychol.*, 1963, 47, 68-71.
- WILKINS, W. L. *The identification of character and behavior disorders in the military life*. Washington, D. C.: United States Department of the Navy, Bureau of Medicine and Surgery, 1961.

(Received February 1, 1963)

COMPENSATORY REACTION TO ANGULARLY DIS- PLACED VISUAL FEEDBACK IN BEHAVIOR¹

CHARLES McDERMID² AND KARL U. SMITH

University of Wisconsin

Compensatory reaction to displaced visual feedback is investigated by requiring S to operationally negate the effects of the feedback displacement by adjusting the direction of response. In this experiment, such compensatory behavior has been compared with direct reactions to the same conditions of angularly displaced vision. As predicted from systematic theory, the efficacy of the compensatory response is generally less than that of direct reaction to displaced vision, both initially and finally in learning. The results also show that learning to respond to displaced visual feedback is specific to the direction and type of compensatory and direct modes of response used. The methods and results of the experiment illustrate how the techniques of displaced and delayed sensory-feedback analysis, using scientific television instrumentation, can be applied to study of the general problems of compensatory behavior, as used in different patterns of adaptive behavior and in response to machine systems of specialized design.

Compensatory reaction to a condition of displaced vision is defined as a reaction in which the direction of movement negates in part the effects of the space displacement and thus the operational effect of the motion appears in a normal direction. This contrasts with direct reaction to visual disorientation, in which the direction of movement and its operational effects are observed as occurring in the direction of the sensory disorientation.

Figure 1 compares the general nature of these two forms of reaction to inverted sensory feedback in writing. In the example shown, the subject is attempting to write the letter "t" while observing all of these movements and their operational effects in a television monitor rather than directly. The circuits of the television camera are modified so that the visual feedback image of motion may be inverted or reversed by the throw of a switch. In one condition of performance, that seen to the left, the subject is instructed to react directly to the inverted feedback, that is, to write so that the movements and

the "t" appear inverted on the monitor. In a second condition, that shown to the right in the figure, the subject is instructed to write in a compensatory way, so that writing of the "t" appears upright in the television monitor even though the visual feedback and the actual movements are inverted.

In a number of experiments extending over several years (Rhule & Smith, 1959; Smith, 1961; Smith & Smith, 1962), utilizing optical and television methods, direct and compensatory reactions to different conditions of reversed and inverted vision have been compared. In these studies the level of performance and degree of adaptation to the displaced sensory feedback have been measured. Because the compensatory pattern of response involves both alteration of established modes of motion as well as displaced visual feedback effects, the assumption underlying the different studies was that the compensatory reaction would show a lower level of performance and a more limited degree of adaptation with practice than the direct pattern of response. It was assumed also that the more complex the motion performed, the sharper the differences between direct and compensatory performance would be. The results of various studies generally confirmed these expectations.

¹ This experiment's funds came from the National Science Foundation (NSF-G-7589) and from the Public Health Service (PHS-4469). This research project enjoys the general collaboration of W. M. Smith, Dartmouth College. The Numerical Analysis Laboratory, the University of Wisconsin, aided in data processing.

² Now in Evanston, Illinois.

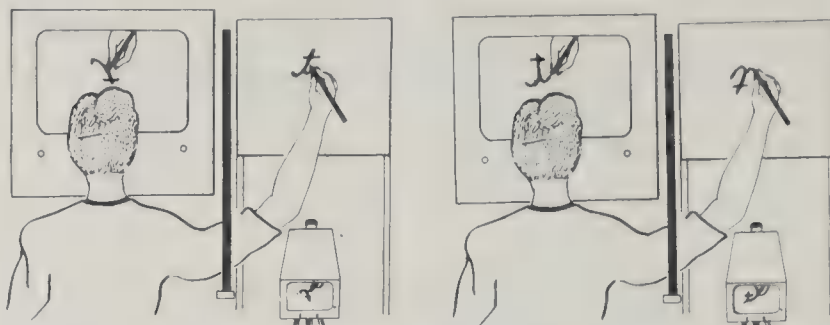


FIG. 1. The difference between direct and compensatory reaction to inverted visual feedback in handwriting. (In direct reaction, shown to the left, the television camera inverts the visual feedback of upright writing so the subject sees this performance as inverted. In compensatory reaction, the television camera inverts the visual feedback of writing, but the subject inverts his movements in writing so that the operational effect of the writing is seen as oriented in an upright position.)

METHOD

In this experiment, the object was to compare the relative effects of direct and compensatory reactions to different conditions of angularly displaced visual feedback in behavior: writing, drawing, and dotting motions. Experimental closed circuit television methods were used in producing the angular displacements. As shown in Figure 2, four different displacements were used: 0, 90, 180, and 270 degrees in the horizontal plane of the performance field. They were produced by moving the camera on its dolly. In all of these conditions, the angle of inclination of the television camera in the vertical plane was comparable to the normal

angle of visual regard in this plane. The general effect of these displacements was that of removing the subject's eyes from his head and, keeping them still functional, placing them in the different angular positions shown.

In all, eight different conditions of performance were used. One was normal vision. A second was zero displacement of the televised feedback. The remaining six came from using conditions of direct and compensatory reaction at each of the other three conditions of angular displacement. In the direct pattern of reaction, the subject's writing, drawing, and dotting movements appeared as angularly displaced from their normal position. In the compensatory reaction, he was required to write "k's" and draw triangles so that they always appeared upright (the triangles with base down) on the television monitor.

Forty-eight right-handed subjects were divided randomly into eight groups of six subjects each, and the specific groups were assigned one of the experimental conditions just described. Each subject practiced the writing, drawing, and dotting tasks for 15 to 30 minutes each day for 9 days.

An electronic handwriting analyzer of our own design and construction was employed to measure separately the contact and travel movements in the writing, drawing, and dot location tasks. This device operates on the principle of using the subject's body as an electrical key. A current of about 100 microamperes is passed through his body and through the pencil with which he writes or draws. Electrical conductive paper is used for the writing and drawing. When the subject touches the paper with his pencil, he closes a relay which operates a precision time clock. The relay is arranged with a flip-flop circuit so that a second clock, which measures the travel time between contacts, is made to operate when contact with the paper is broken in the travel movement and the first clock stops.

The subject wrote "k's," drew triangles, and made dots in a sequential manner. When he reached the edge of the paper he touched a small stop plate

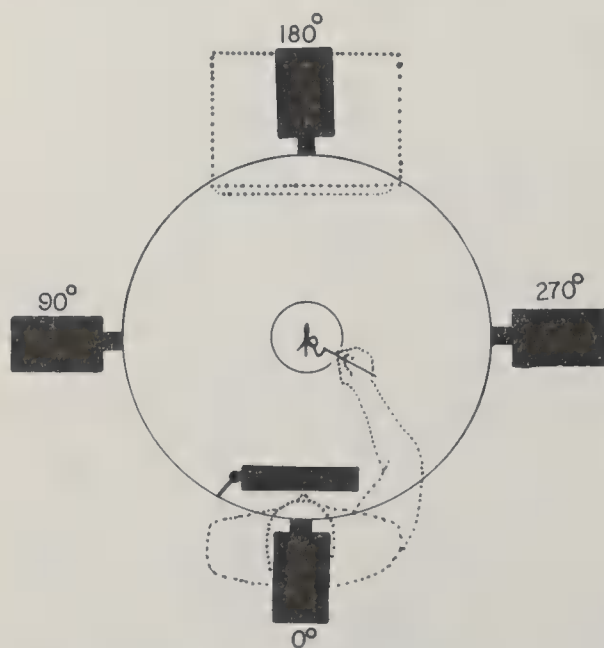


FIG. 2. Conditions of angular displacement of the visual feedback of writing, drawing, and dot location motions used in this study.

which automatically stopped the timing in a particular trial. Both contact time and travel time were recorded in the three different tasks of writing, drawing, and dot location.

The assumptions under test in this experiment were that: (a) the level of performance, especially initially in adaptation, will favor the direct mode of reaction because it involves fewer dimensions of relative visuomotor displacement; (b) the nature of the learning functions with direct and compensatory modes of response will reflect specifically the general effects of the relative displacement of vision and motion.

RESULTS

The learning curves given in Figure 3 summarize the details of the results concerning total time of performance on successive days of practice under the eight different experimental conditions. The graph plots motion time in seconds as a function of practice days. The letter "D" with the displacement angle value designates direct

reaction and the letter "C" compensatory reaction.

Three main effects are indicated. Time of performance varied greatly as a function of relative angular displacement of both visual feedback and motion. The form and level of the learning function was determined specifically by the magnitude and nature of relative angular displacement of visual feedback and movement. The adverse effects of the relative displacement were greatest both initially and finally in practice for the compensatory type of reaction. Variance analysis and Duncan (1955) range tests indicated that on Day 9 of practice the 0-, 90-, and 180-degree compensatory displacements were significantly different from most other conditions. The 90-degree direct displacement also varied significantly from most other conditions. These were the conditions that produced the most severe effects on performance initially.

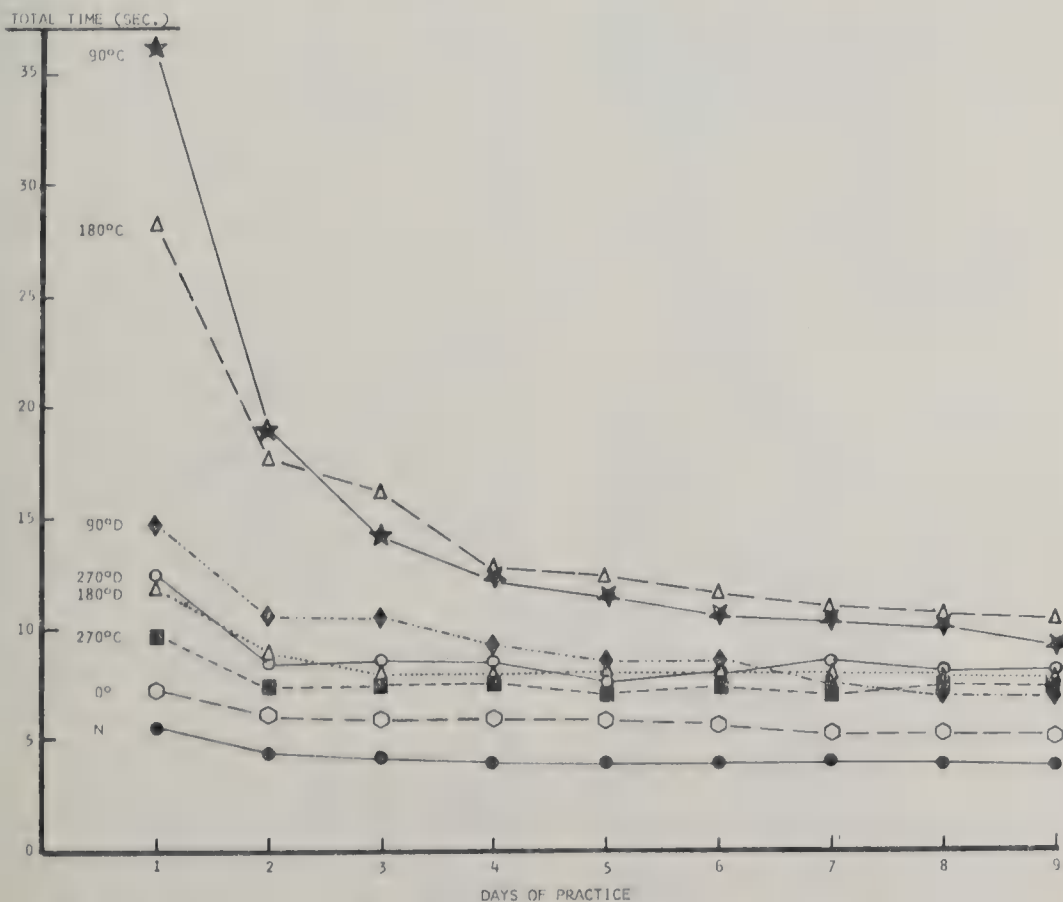


FIG. 3. Learning curves indicating the differences between performance at different displacement conditions throughout the 9 days of practice. (The letter "C" alongside a particular displacement value denotes compensatory reaction and a letter "D" denotes direct reaction; the letter "N" refers to normal vision.)

Some of the main details of the quantitative differences between displacement conditions, as found in this study, are summarized in Tables 1 and 2. Table 1 gives figures covering all 9 days of performance while Table 2 presents separate analyses for performances on Days 1 and 9. The data in these tables represent the results of Duncan multiple-range tests of the significance of differences between means for different displacement conditions. In Table 1, separate results are given for contact and travel time means for the three tasks. In Table 2, separate figures are given for contact and travel means on Days 1 and 9; in this table the data for all performances are combined. The columns in the two tables list in order the magnitudes of the means for different displacement conditions. Common bars alongside any two or more displacement conditions denote that these conditions are not statistically significant from one another at the 1% level.

Examination of the two tables will show that there is considerable consistency in the results. As stated earlier, the compensatory modes of reaction generally gave the longest

TABLE 1
DUNCAN MULTIPLE-RANGE TESTS OF THE SIGNIFICANCE OF THE DIFFERENCES BETWEEN DISPLACEMENT CONDITIONS

	Dot location	Drawing triangles	Writing k's
Contact time	ns	90C	90C
		180C	180C
		90D	90D
		270D	270D
		180D	180D
		270C	270C
		0	0
		N	N
Travel time	ns	180C	180C
		90C	90C
		90D	180D
		180D	270C
		270D	90D
		270C	270D
		0	0
		N	N

Note.—Any two conditions displaying a side bar in common are not significantly different from one another. The data in this table are based on *F* values significant at the 1% level.

TABLE 2
DUNCAN MULTIPLE-RANGE TESTS OF THE SIGNIFICANCE OF THE DIFFERENCES BETWEEN DISPLACEMENT CONDITIONS ON DAY 1 AND DAY 9

Contact time		Travel time	
Day 1	Day 9	Day 1	Day 9
90C	90C	90C	180C
180C	180C	180C	180D
90D	270D	90D	90C
270D	180D	180D	270C
180D	90D	270D	270D
270C	270C	270C	90D
0	0	0	0
N	N	N	N

Note.—Any two conditions displaying a side bar in common are not significantly different from one another. The data in this table are based on *F* values significant at the 1% level.

mean durations in performance. The 90-degree and 180-degree displacement conditions were generally more difficult than the 270-degree. In general, the 90-degree and 180-degree compensatory conditions are set off as significantly different from other conditions throughout practice and on both the first and last day. The two zero conditions of displacement are also generally set off as significantly different from other conditions except for contact and travel time means on Day 1.

The results are generally clear in showing that marked differences occur with variation in magnitude of angular displacement of vision and with variation in direction of movement under a given condition of displacement. The only condition which appears out of some expected order is the 270-degree condition. However, this angle of displacement is not to be considered in terms of the absolute value, but as a 90-degree displacement to the right of the subject. Inasmuch as all of the subjects used were right-handed, this angle is much easier for performance in writing than the same magnitude of displacement to the left of the performance field. These results are in keeping with other findings that handedness is related to angular displacement of sensory feedback control of movements.

As shown in Figure 3 the nature of the learning functions varies specifically as a function of displacement conditions and mode

of response. The more difficult conditions produce initially the most inaccurate and slowest reactions, the greatest absolute amount of learning and the most limited final level of adaptation. The results found are specifically in keeping with the original assumption that the level and rate of learning are explicit derivative effects of the relative geometricity of displacement conditions. All the different angular displacements gave significant learning effects.

Performing compensatory reactions to displaced vision involves the adjustment of direction of motion in order to correct, in part, the perceptual effects of the displacement. This corrective perceptual effect, however, is not complete because it is mainly the operational aspect of the motion that is affected, such as the direction of writing or drawing or the positioning of an object. The movements actually performed with the displacement are still observed dynamically in the direction determined by the displacement. Thus, the overall effect of this is to aid perceptual orientation of the subject in the inverted or angularly displaced visual field. In prior studies (Smith, 1961), it has been found that if compensatory types of reaction are performed some subjects never become aware that their visual world is inverted during the experiment. A similar result was found in this study with angularly displaced visual feedback in that subjects in different compensatory groups did not know that the visual orientation of their movements were in fact displaced. These results indicate a possible source of the fleeting and arbitrary judgments of orientation of the perceptual world that subjects wearing experimental spectacles are sometimes reported to make (Ewert, 1930; Kohler, 1955; Snyder & Pronko, 1952; Stratton, 1896, 1897, 1899).

DISCUSSION

Compensatory reaction to displaced sensory feedback represents a general mode of behavior that is observed in many situations of human skill and machine performance, such as in compensatory tracking, corrective steering, corrective positioning movements, postural balancing reactions, and many types of

fixation movements of the eyes and other members of the body. The study of such behavior is also significant because compensatory reaction to inverted vision has been unknowingly involved in past studies of inverted visual perception, utilizing experimental spectacles.

This experiment's results may be described generally as follows:

1. Compensatory reaction to displaced vision is one method of response with which the subject can partially correct for the disorientation produced by inversion, reversal, angular rotation, or angular displacement of vision.

2. As predicted from systematic theory based on sensory feedback conceptions of displaced perception and motion, efficiency of performance and degree of adaptation with compensatory reactions are more limited than with direct reactions to angular displacement of vision.

3. The relative effects of different magnitudes of angular displacement of the locus of vision vary in much the same way for the direct and compensatory modes of adaptation. The degree of disorientation of movement increases with the magnitude of displacement. However, these quantitative variations vary in relation to both the axis or plane of displacement of the visual feedback and the subjects' characteristics of handedness.

4. The general result of this study is that the effects of displaced sensory feedback in behavior are not based upon absolute orientation of the sensory or motor systems in space, or upon built-in stimulus-response correspondences. Rather, the effects of different conditions of inversion, reversal, and angular displacement of vision feedback depend on the relative geometric orientation and displacement of vision and movement.

5. There is definite evidence from this study that both direct and compensatory reactions to different magnitudes of angular displacement of vision are learned as specific patterns of behavior. In none of the displacement conditions studied did the subjects reach normal levels of performance after daily periods of practice extending over 9 days.

6. Compensatory reaction to displaced sensory feedback appears to be a major means

whereby the motor system controls the properties of perceptual orientation and organization, and determines the pattern of overt response relative to such response produced perceptual organization.

7. Compensatory reactions to inverted vision apparently have been involved in all prior studies of perceptual adaptation using experimental spectacles which invert the visual field and have influenced the variable judgments about perceptual orientation reported in these studies.

8. The direct and compensatory modes of response are not the only ways of adjusting to displaced sensory feedback. In addition, the subject can be required to orient his motions in any one of hundreds of discriminable directions and planes of motion with any one particular axis or magnitude of sensory displacement. Accordingly, the present study defines one aspect of a broad spectrum of phenomena related to the spatial interaction of the sensory and motor systems in behavior which may have significance for understanding the determination of organization of behavior in man.

REFERENCES

- DUNCAN, D. B. Multiple-range and multiple F tests. *Biometrics*, 1955, **11**, 1-42.
- EWERT, P. H. A study of the effect of inverted, retinal stimulation upon spatially coordinated behavior. *Genet. Psychol. Monogr.*, 1930, **7**, 177-363.
- KOHLER, I. Experiments with prolonged optical distortion. *Acta psychol., Amsterdam*, 1955, **11**, 176-178.
- RHULE, W., & SMITH, K. U. Effects of inversion of the visual field on human motions. *J. exp. Psychol.*, 1959, **57**, 338-343.
- SMITH, K. U. The geometry of motion and its neural foundations: II. Neurogeometric theory and its experimental basis. *Amer. J. phys. Med.*, 1961, **40**, 109-129.
- SMITH, K. U., & SMITH, W. M. *Perception and motion: An analysis of space structured behavior*. Philadelphia, Pa.: Saunders, 1962.
- SNYDER, F. W., & PRONKO, N. H. Vision with spatial inversion. Wichita: Univer. Wichita Press, 1952.
- STRATTON, G. M. Some preliminary experiments in vision without inversion of the retinal image. *Psychol. Rev.*, 1896, **3**, 611-617.
- STRATTON, G. M. Vision without inversion of the retinal image. *Psychol. Rev.*, 1897, **4**, 341-360, 463-481.
- STRATTON, G. M. The spatial harmony of touch and sight. *Mind*, 1899, **8**, 492-505.

(Received February 14, 1963)

CULTURAL VALUES AND EMPLOYEE ATTITUDES: UNITED STATES AND JAPAN

ARTHUR M. WHITEHILL, JR.

University of North Carolina

Often overlooked in industrial motivation studies is the influence of cultural orientation upon employee attitudes. This is particularly true with respect to the way workers feel about reciprocal obligations in employee-employer relations. Attitudes concerning a broad range of such obligations are influenced strongly by the cultural values indigenous to a given society. For example, such values have been shown to influence worker perception of such matters as employment continuity, economic involvement and personal involvement of management, identification with organization, status transfer, sources of motivation, and other aspects of the work situation. Mutual satisfaction in human relations in industry depends partially upon management understanding of and willingness to work through cultural values and the employee attitudes they engender.

Studies of worker motivation traditionally have stressed such elements as organizational structure, nature of supervision, social relations and, of course, wages and other forms of recognition. Many situations remain, however, in which these elements are "favorably" represented, yet motivation remains low. Conversely, there are other situations in which these elements appear to be distinctly "unfavorable," yet motivation obviously is extremely high.

An example of the latter, from the viewpoint of Western observers, is the amazing will to work demonstrated by individual Japanese employees *without the support* of incentives, personnel policies, and patterns of leadership considered essential in the United States to high levels of motivation. The research program reported upon here was conducted in Japan and the United States. Its findings shed new light upon an important, though often ignored, dimension of motivation theory: the pervasive impact of cultural values upon worker attitudes and behavior.

In a sense, the study has provided a laboratory demonstration of the validity of a statement by the director of the Motivations Project at Massachusetts Institute of Technology, that:

Interpersonal relationships which will be effective in economic activity in a given country depend on the country's culture. Principles of business administration are not absolute; they are relative to the society [Hagen, 1958, p. vii].

Particularly sensitive to cultural orientation is the way employees feel about reciprocal obligations—the *quid pro quo* or "something for something"—which to varying degrees and with multiple meanings characterize worker-management relations throughout the industrialized world. The present article deals with this reciprocity in such matters as employment continuity, economic and personal involvement, identification with organization, and status transfer. In addition, a note is added concerning cross-cultural differences in worker perception of compelling sources of motivational pressures.

Before giving some relevant findings from the research data, the usage of two terms employed throughout the remainder of this article should be made clear. Reference to "culture" is not made in any esoteric sense but, rather, to the whole complex of distinctive features characteristic of a particular stage of advancement in a given society. The term "value" is used in the sense of "a selective orientation toward experience, implying deep commitment or repudiation, which influences the ordering of 'choices' between possible alternatives in action [Kluckhohn, 1961, p. 18]."

CULTURAL VALUE COMPARISONS

In an attempt to investigate systematically the impact of cultural values upon worker and management behavior, the investigator

and a Japanese colleague, Shin-ichi Takezawa of Rikkyo University in Tokyo, initiated what has developed into a rather ambitious, cross-cultural study. Thus far, a survey of 2,000 production workers, equally divided between Japan and the United States, and employed by four roughly comparable firms in each of the two countries has been completed. Plans call for extension of the study to other south-east Asian countries during the next few years.

Theoretical Framework

The research is based upon an evolving concept which might be called a "theory of reciprocal role expectations." A basic proposition underlying the study is that how a man thinks and acts in a given situation will be significantly influenced by what he thinks is appropriate, proper, and expected of him. However, whether or not his behavior coincides with this role expectation will depend to a considerable extent upon how faithfully those with whom he interacts fulfill *his* conception of their proper role; that is, how "fair" their behavior seems to him.

In short, behavioral decisions are made at least partially on the basis of what we feel is expected of us and how well others are fulfilling our expectations of them. Here again, some dimensions of the reciprocal relationship referred to earlier may be seen.

A final and extremely important point in this theoretical schema is that the nature of these expectations will be molded by the total environment within which this mutual evaluation takes place *and will vary markedly from one culture to another!* Therefore, there is always a need for understanding the cultural forces at work if negotiations are to be successful for whatever parties are involved. Several examples drawn from the study conducted in Japan and the United States may clarify this point of view.

Employment Continuity

On the basic issue of employment continuity, workers throughout the world have indicated a strong desire for what seems to them "reasonable" employment security. But what actually constitutes "reasonableness" in this context varies substantially from culture

to culture. Hence, the expectations of workers concerning employment continuity will also vary markedly among different industrial societies.

A relevant question asked in the survey of workers in United States and Japanese firms follows, with the percentage responses from the two groups.

If a worker, although willing, proves to be unqualified on his job, management should feel a responsibility:	United States (%)	Japan (%)
1. to continue his employment until he retires or dies;	23	55
2. to continue his employment for as long as 1 year so that he may look for another job;	19	23
3. to continue his employment for 3 months so that he may look for another job;	38	18
4. to terminate the employment of unqualified workers after giving about 2 weeks notice.	20	4

It is quite clear that the expectations of Japanese workers concerning the role of "good management" in providing employment continuity are considerably more exacting than their American counterparts. Even though a worker proves to be unsatisfactory, more than half the Japanese respondents felt that management should continue his employment indefinitely.

Economic Involvement

Turning for a moment to attitudes concerning economic involvement in aspects of employees' lives not directly related to their work, considerable divergence was found among respondents when questioned concerning the provision of housing assistance by management.

In regard to housing for workers, management should:	United States (%)	Japan (%)
1. provide company housing at no charge;	2	29
2. provide company housing at special low rent;	8	39
3. provide low-interest loans to assist workers in owning their own homes;	56	29
4. avoid direct, financial assistance in housing.	34	3

Here we see that about two thirds of Japanese respondents, as compared with only one tenth of the United States participants, felt that a well-managed company should own housing facilities and make them available to workers on a no charge or low rent basis. Obviously, managers anxious to fulfill this role expectation of workers in their respective cultural settings must proceed quite differently as things now stand.

Personal Involvement

For a final example—many others could be cited from the research findings—let us take a strictly personal matter such as marriage and see the intercultural range in worker’s attitudes.

When a worker wishes to marry, I think his (her) supervisor should:	United States (%)	Japan (%)
1. help select a possible mate and serve as go-between;	2	6
2. offer personal advice to the worker if requested;	29	70
3. merely present a small gift from the company;	9	19
4. not be involved in such a personal matter.	60	5

Again we see that a manager, in fulfilling the expectations of his workers, proceeds at his own peril if he does not fully understand the cultural complex within which these expectations are formulated.

Now to carry one step further the reciprocity in employee-employer relations being explored here, attention should be directed to the impact of cultural forces upon the responses to several questions which explore the other side of the coin; that is, the obligations workers feel toward management in return for recognizing their expectations in these human relations matters.

Identification with Organization

Managers throughout the world seem to seek from employees a demonstration of loyalty in the sense of a positive identification with the company. This, in turn, demands some transference of the worker’s emphasis upon his personal life to a greater acceptance

of the importance of his on-the-job life. The following question indicates that the extent to which workers are willing to fuse their personal and job lives to meet this management goal varies substantially among industrial societies.

I think of my company as:	United States (%)	Japan (%)
1. the central concern in my life and of greater importance than my personal life;	1	9
2. a part of my life at least equal in importance to my personal life;	23	57
3. a place for me to work with management, during work hours, to accomplish mutual goals;	54	26
4. strictly a place to work and entirely separate from my personal life.	23	6

There seems to be little doubt that Japanese participants are more willing to identify themselves positively with the company than are United States respondents. More than two thirds of the Japanese workers accorded their job lives at least equal importance when compared with their personal lives. The response pattern was almost reversed by United States participants.

Status Transfer

Along the same line, but in a lighter vein, a question was designed to explore the attitudes of workers in the two societies concerning transfer to off-job situations of the status distinctions found in a company’s formal organization. The assumption here is that greater willingness in this respect is indicative of closer identification of the individual with the company and its goals.

If my immediate supervisor enters a crowded bus on which I am riding I should:	United States (%)	Japan (%)
1. always offer him my seat since he is my superior;	2	10
2. offer him my seat unless I am not feeling well;	2	44
3. remain seated and offer to hold any packages he may have;	33	41
4. remain seated since a fair rule is “first come, first served.”	63	5

Motivational Sources

Finally, there are more subtle distinctions, too, that should be understood by managers if they are to respond in a meaningful fashion to the feelings and reactions of workers. Differences in attitudes concerning the sources of motivation to perform “a fair day’s work for a fair day’s pay” is a case at point. What sorts of responsibilities do employees feel for performing well on the job? How important are wages, for example, as a source of worker motivation? The following question reveals some significant differences in these matters among United States and Japanese respondents.

I believe workers are willing to work hard on their jobs because:	United States (%)	Japan (%)
1. they want to live up to the expectations of their family, friends, and society;	10	41
2. they feel it is their responsibility to the company and to co-workers to do whatever work is assigned to them;	61	37
3. the harder they work, the more likely they are to be promoted over others to positions of greater responsibility;	9	11
4. the harder they work, the more money they expect to earn.	20	11

Comparison of Items 1 and 2 shows a welcome assumption of a sense of responsibility among workers in both cultures. But the response pattern also clearly indicates the broader, more generalized base upon which this quality is founded among Japanese, as compared with United States employees. Demonstrated less dramatically is the somewhat greater emphasis placed by United States workers upon money as a motivating force.

CONCLUSIONS

It would be risky to formulate sweeping generalizations on the basis of the limited research findings reported here. Yet, the following conclusions seem to be persuasive on at least a tentative basis.

First, reciprocity, in the sense of willingness to enter into a set of reciprocal obligations, is perceived by workers to be an appropriate aspect of their relationship with employers.

Second, cultural forces indigenous to a given society tend to mold the attitudes of workers as to what they may reasonably expect in the human relations area from good management. Furthermore, the reciprocal obligations which they as good workers are willing to extend to management are also often culturebound.

Third, in exercising its position of leadership, management success in human relations will depend, at least in part, upon understanding *in depth* the nature and impact of the cultural environment which exerts a significant influence upon worker attitudes and behavior.

A final conviction arising from the research program is that organizational leaders must realize that the forces affecting man’s willingness to work are dynamic in nature and *can* be influenced by creative, alert administrators. In so doing, it will be necessary to explore anew prevailing cultural values which create barriers against, and promising possibilities for, reciprocity—in the sense of offering something for something—in the human, as well as the technical and legal aspects of administration.

REFERENCES

HAGEN, E. E. Foreword to J. G. Abegglen. In, *The Japanese factory*. Glencoe, Ill.: Free Press, 1958.
KLUCKHOHN, C. The study of values. In D. N. Barrett (Ed.), *Values in America*. Notre Dame: Univer. Notre Dame Press, 1961.

(Received February 15, 1963)

Journal of Applied Psychology

KENNETH E. CLARK, Editor
University of Rochester

Table of Contents

Edward Kellogg Strong, Jr.: 1884-1963.....	John G. Darley	73
Proposed Scoring Changes for the Strong Vocational Interest Blank..... :.....E. K. Strong, Jr., David P. Campbell, Ralph F. Berdie, and Kenneth E. Clark		75
Relation between Radar Detection and the Observer's Concept of a Target.....R. D. Baldwin, A. D. Wright, and D. J. Lehr		81
Application of a Method of Evaluating Training.....	John A. Cox	84
Sex Differences in Job Satisfaction.....	Charles L. Hulin and Patricia Cain Smith	88
Brand Loyalty Revisited: A Twenty-Year Report.....	Lester Guest	93
Test Difficulty, Reliability, and Discrimination as Functions of Item Difficulty Order.....	Marshall H. Brenner	98
Life History Antecedents of Sales, Research, and General Engineering Interest.....	Frederick B. Chaney and William A. Owens	101
Supervisor Esteem and Personnel Evaluations.....	Paul D. Nelson	106
Comparison of Programed and Conventional Instruction Methods.....	Myles H. Goldberg, Robert I. Dawson, and Richard S. Barrett	110
Tracking with a Differential Brightness Display: I. Acquisition and Transfer.....	Stanley M. Moss	115
Characteristics of Successful Policemen and Firemen Applicants.....	Joseph D. Matarazzo, Bernadene V. Allen, George Saslow, and Arthur Wiens	123
Predictor Variables for Creativity in Industrial Science.....	Francis E. Jones	134

American Psychological Association

Consulting Editors

GEORGE E. BRIGGS, *Ohio State University*
NORMAN FREDERIKSEN, *Educational Testing Service*
LEONARD D. GOODSTEIN, *University of Iowa*
EDWIN R. HENRY, *Peace Corps*
JOHN HOLLAND, *American College Testing Program*
CLIFFORD E. JURGENSEN, *Minneapolis Gas Company*

LAURENCE S. MCGAUGHRAN, *University of Houston*
QUINN MCNEMAR, *Stanford University*
HAROLD F. ROTHE, *Beloit Corporation*
THOMAS A. RYAN, *Cornell University*
ALEXANDER G. WESMAN, *Psychological Corporation*
CLARK L. WILSON, *Batten, Barton, Durstine, and Osborn, Incorporated*

This journal gives primary consideration to original investigations in any field of applied psychology except clinical psychology, although a descriptive or theoretical article may be accepted if it represents a special contribution in an applied field. Quantitative investigations of interest or value to psychologists working in the following broad fields will be considered: vocational and educational prognosis, diagnosis, and guidance at the secondary and college level; personnel research in business, industry, and government; engineering psychology; industrial working conditions; research on opinion and morale factors; job analysis and classification research; market and advertising research.

Manuscripts must be accompanied by an abstract of 100–120 words typed on a separate sheet of paper. The abstract should conform to the style of *Psychological Abstracts*. Detailed instructions for preparation of the abstracts appeared in the *American Psychologist* (1961, 16, 833), or they may be obtained from the Editor or from the APA Central Office.

Manuscripts should be addressed to the Editor:

Dr. Kenneth E. Clark
College of Arts and Science
302 Morey Hall
University of Rochester
Rochester, N. Y. 14627

All manuscripts must be submitted in duplicate. Original figures are prepared for publication; duplicate figures may be photographs or pencil-drawn copies.

Manuscripts must conform to the style requirements described in the *Publication Manual of the American Psychological Association*.

The following policies govern reprinting of materials copyrighted in APA journals: (a) to require approval to reprint for tables and figures, and for text only if more than 500 words [total from one article] in length; (b) to continue approval to reprint, contingent on the author's approval, for articles reprinted in whole or in major part; (c) to negotiate where possible the dedication of royalties from commercial publishers to the American Psychological Foundation.

Journal of Applied Psychology

Published bimonthly by the
American Psychological Association
Prince and Lemon Sts., Lancaster, Pa. 17604
and 1333 Sixteenth St. N. W.
Washington, D. C. 20036

\$10.00 per volume

\$2.00 per issue

HELEN ORR
Managing Editor

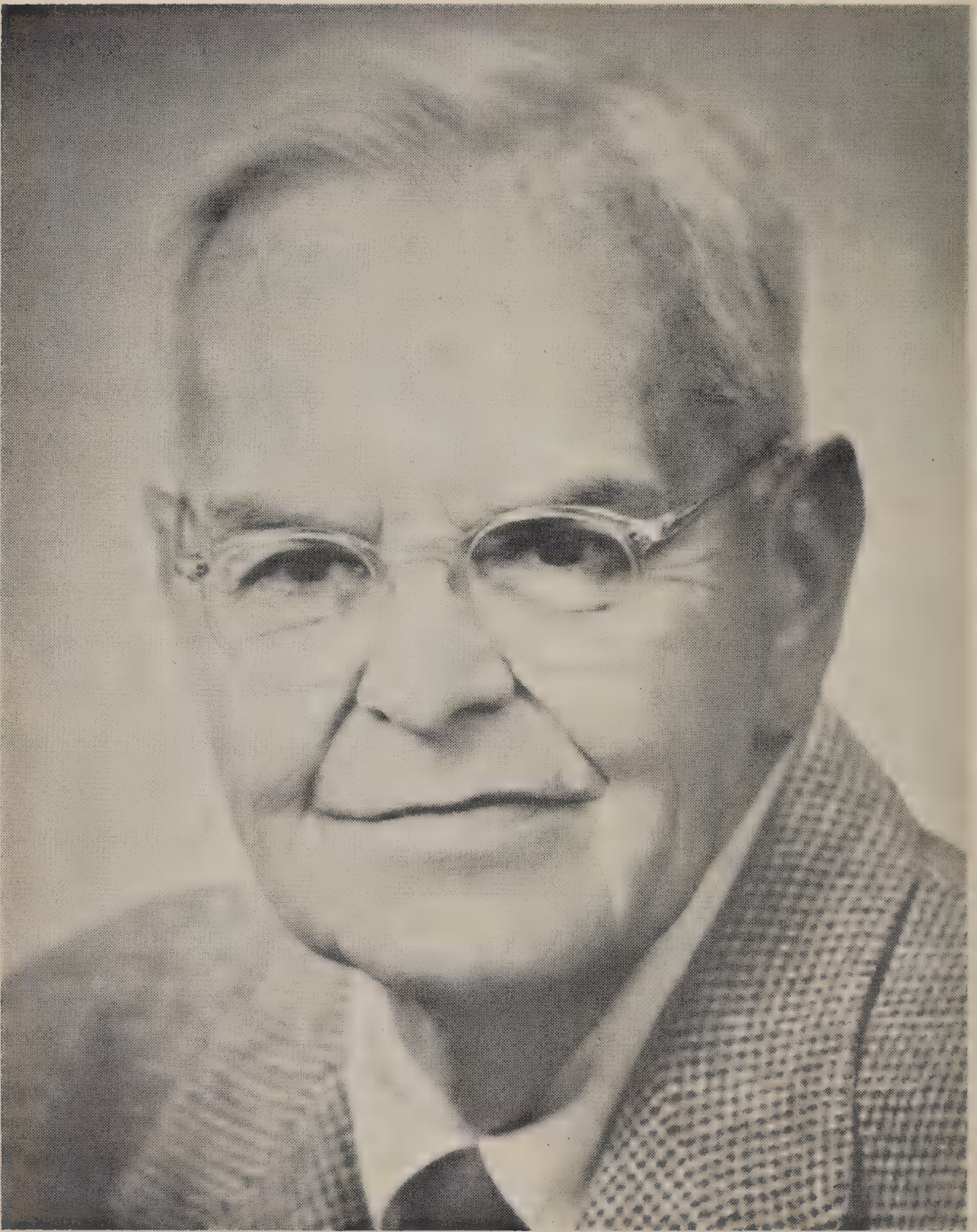
ELIZABETH S. REED
Advertising Manager

FRANCIS L. BREWER
Editorial Assistant

Subscriptions, orders, and business communications should be addressed to the American Psychological Association, 1333 Sixteenth St. N. W., Washington, D. C. 20036. Address changes must reach the subscription office by the tenth of the month to take effect the following month. Undelivered copies resulting from address changes will not be replaced; subscribers should notify the post office that they will guarantee second-class forwarding postage. Other claims for undelivered copies must be made within four months of publication.

Second class postage paid at Lancaster, Pa. and at additional mailing offices.

© 1964 by the American Psychological Association, Inc.



EDWARD KELLOGG STRONG, JR.

Journal of Applied Psychology

VOL. 48, No. 2

APRIL 1964

EDWARD KELLOGG STRONG, JR.

1884-1963

Edward K. Strong, Jr., Emeritus Professor of Applied Psychology, Stanford University, died on the night of December 4, 1963, Menlo Park, California. Thus was broken another physical link in one of psychology's two great chains of empiric research—the field of vocational interest measurement. Stanford University was also the locus of psychology's other long-continuing research tradition, in Terman's work on the measurement of intelligence. It was Terman who in 1923 brought Strong to the Stanford faculty. In the years that followed these two research traditions provided a large part of the knowledge from which psychology speaks today to the questions of what man can do—his ability—and what man wants to do—his interests. In the *Genetic Studies of Genius* and the *Vocational Interests of Men and Women* will be found the record of the pioneering explorations of these two questions.

Dr. Strong was born in Syracuse, New York, in August of 1884. When his family moved to the pastorate of a Presbyterian church in San Francisco in 1902, he entered the University of California, graduating in 1906 with a major in biology. After a short period in the United States Forestry Service, he returned to Berkeley for his Master's degree in psychology (1909) and then went on to Columbia, receiving his PhD in 1911. There James McKeen Cattell, "probably the most influential of the second generation" of scientific psychologists, must have helped shape Strong to his career in applied psychology, which then started with the next 3 years spent in advertising, marketing, and business research at Columbia.

From 1914 to 1917, Dr. Strong was Professor of Psychology at George Peabody College for Teachers. In 1917 he entered military service, serving first on the famous Committee on Classification of Personnel and ultimately holding military rank on the Army General Staff; he served as the personnel specialist at Camp Taylor, Kentucky, and Camp Kearney, California.

From 1919 to 1923, Dr. Strong was a member of the faculty of the Carnegie Institute of Technology, the time and the place in which applied psychology had a period of its most concentrated growth. Here the work on interest measurement began; Dr. Strong acknowledges, in the preface to his own book, his debt at Carnegie to the work of J. B. Miner, C. S. Yoakum, Max Freyd, and, in the early days at Stanford, Karl Cowdery. From 1923 onward, coinciding with his appointment at Stanford, his consuming research concern was the field of interest measurement. The first edition of the Strong Vocational Interest Blank was published in 1927; the revised edition in 1938; the major volume, *Vocational Interests of Men and Women*, in 1943; the Stanford follow-up study, *Vocational Interests Eighteen Years After College*, in 1955. These were the landmarks; there were, in addition, many articles in the technical journals on the same basic topic.

Dr. Strong's death brings a particular sense of loss for many of us at Minnesota, for we knew him over the years as friend and frequent visitor and colleague. The long friendship between Dr. Strong and the late Donald G. Paterson caught Minnesota up in the orbit of Dr. Strong's dominant concern, and many of us have spent some part of our professional careers in the exciting domain of interest measurement and theory. A recent report on research in this area lists at least 58 theses, 46 technical journal articles, and 6 monographs or books produced at Minnesota. The University has also established a Center for Interest Measurement Research, under the direction of Professor David P. Campbell; to this Center Dr. Strong, before his death, had given all his research files and materials on the vocational interest inventories. Dr. Strong had delivered the 1958 Walter Van Dyke Bingham lecture at Minnesota, honoring both the speaker and the institution for their dedication to this aspect of Dr. Bingham's concern for research in human talent.

As Dr. Strong has said, "The study of interests

was initiated in an atmosphere of applied psychology. Most of the worthwhile work has been directed toward the use of interests as a means of solving practical problems." It is impossible to estimate how many thousands of young people have been helped in crucial career choices by the use of the Strong Vocational Interest Blank. But its contribution to men's success and satisfaction is incalculable, and is in the great tradition of American applied psychology. In recent years its value as a research source in the areas of personality theory and motivation has begun to emerge. Man's working life—reflecting in substantial part his needs, his life styles, his motivations, and his satisfactions—is truly the arena in which theories of personality and motivation must ultimately receive rigorous

testing and verification, since work and careers loom so large in our culture. When task orientation comes under study, the rich store of data from Strong's years of research will be most valuable.

Psychology has yet to weave the theoretical and empirical strands of its own history as a science into the whole cloth of an understanding of human behavior. When this time comes, Professor Strong's empiric contributions to an understanding of man's satisfactions in the world of work will provide the solid foundation of data without which theory cannot be built. His gifts to psychology mitigate the sense of loss that his death brings to his many friends and colleagues.

JOHN G. DARLEY

PROPOSED SCORING CHANGES FOR THE STRONG VOCATIONAL INTEREST BLANK¹

E. K. STRONG, JR.
Stanford University

RALPH F. BERDIE
University of Minnesota

AND

DAVID P. CAMPBELL
University of Minnesota

KENNETH E. CLARK
University of Rochester

A study of the effect of replacing with new items 102 items in the current Strong Vocational Interest Blank (SVIB) on the validity and reliability of the SVIB. 8 occupational groups were used. The results showed that the validity and reliability remained essentially the same after dropping the items. Also, the weighting system of the SVIB was compared with scales using unit weights. The results indicated that the unit weights keys were virtually identical to the weighted keys on 3 criteria: validity, reliability, and scale intercorrelations. Thus, the SVIB when revised should be scored using unit weight scales.

In an earlier article, Strong has discussed the need for revision of the Strong Vocational Interest Blank (SVIB), and listed some contemplated changes (Strong, 1962). Specifically, invalid and obsolete items (such as the names of magazines now out of circulation) will be replaced with new items. Of the current 400 items, 102 will be replaced, and 5 new ones will be added for a new total of 405.² These changes apply to the Men's Form only. Obviously this will affect the scoring of the SVIB as the current scales and norms will be unuseable. Until validation data can be collected on these new items, the SVIB will be scored on the 298 unchanged items only.

Will it still be valid using only these unchanged items? To determine this, Strong's original standardization data for eight occupational groups were reanalyzed using only these 298 items. These results were compared with the original results to determine the effect on validity of using this shortened item pool. This report presents those comparisons.

At the same time, the original data were reanalyzed to decide what system of weighting should be used on the revised SVIB. When

the SVIB first came out, the item weights ranged from 30 to -30. In the 1930 revision, these were reduced to 15 to -15. The 1938 revision reduced them further to 4 to -4 where they remain today. Clark's (1961) research indicated that these weights might be dropped even further to 1 to -1, and another recent study (Fruchter, 1963) also supports unit weights over a more complex weighting of items.

On the other hand, Kuder recently has proposed that item weights be established by subtracting the response percentage of one group from that of another group, in effect creating a weighting system of 100 to -100, though in practice it would seldom range that widely. Using this system, he reports extremely low overlaps, but as he did not report the analogous overlaps using unit weights, it is difficult from his report to determine if this complex system is really superior to unit weights (Kuder, 1963).

METHOD

Several different methods of keying were tried. The methods varied in the weights used and the items included in the scale. The weights used were 4 to -4, 3 to -3, 2 to -2, 1 and -1, 4 to 1, 3 to 1, 2 to 1, 1, and -1; the items were selected on the basis of the size of the weight on the original SVIB scale. Table 1 shows the composition of the various keys used.

As an illustration, Key Number 10 used the items currently weighted +4, +3, and +2, weighting them all +1, and used the items now weighted -4, -3, -2, weighting them -1.

¹This study was in part supported by funds from the Graduate School of the University of Minnesota. Computer time was furnished by the Numerical Analysis Center at the University of Minnesota.

²The 5 items will be added to help administrators and scoring services discriminate between the current and the revised booklets and answer sheets.

TABLE 1
COMPOSITION OF EXPERIMENTAL KEYS

Experi- mental keys	Current item weights							
	+4	+3	+2	+1	-1	-2	-3	-4
1 ^a	+4	+3	+2	+1	-1	-2	-3	-4
2	+3 ^b	+2	+1	—	—	-1	-2	-3
3	+2	+1	—	—	—	—	-1	-2
4	+1	—	—	—	—	—	—	-1
5	+4	+3	+2	+1	—	—	—	—
6	+3	+2	+1	—	—	—	—	—
7	+2	+1	—	—	—	—	—	—
8	+1	—	—	—	—	—	—	—
9	+1	+1	+1	+1	-1	-1	-1	-1
10	+1	+1	+1	—	—	-1	-1	-1
11	+1	+1	—	—	—	—	-1	-1
12	+1	—	—	—	—	—	—	-1
13	+1	+1	+1	+1	—	—	—	—
14	+1	+1	+1	—	—	—	—	—
15	+1	+1	—	—	—	—	—	—
16	+1	—	—	—	—	—	—	—
17	—	—	—	—	-1	-1	-1	-1
18	—	—	—	—	—	-1	-1	-1
19	—	—	—	—	—	—	-1	-1
20	—	—	—	—	—	—	—	-1

^a Current key.
^b Item weight on experimental keys.

(A complete item analysis using the original response percentages was not done, as the method described above is a simple way of establishing a varying set of weights highly correlated with weights that might be derived from the response percentages.) All of these experimental keys were applied first to the entire pool of 400 items, then to the reduced pool of 298 items, using the eight occupational groups listed in Table 2.

Keys with ± 4 weights only (Numbers 8, 16, 20) had too few items—usually less than 10, always less than 25—to be considered further. Keys containing only negatively weighted items (Numbers 17–20) were eliminated at this point also as they were less valid than the other methods.

Three kinds of data were used to evaluate the different methods of keying: First, validity data comparing overlap between the occupational groups and “men in general” (MIG); second, test-retest reliability over a 30-day period; and third, inter-correlations between the occupational scales.

RESULTS AND DISCUSSION

The results comparing keys using the entire item pool of 400 items with the same keys using only the unchanged 298 items are presented in Table 2 for both the current key and the “best unit weight” key. The latter is that unit weight key with the lowest percentage overlap and a minimum number of approximately 40 items. Clearly the differences are small. Reducing the item pool will not affect validity. The number of items on the blank is not being reduced in spite of this in order to retain an item pool of

sufficient size and diversity for the development of new interest scales.

In Table 3 are listed the comparisons between weighted keys and unit weight keys. For each occupational scale, the results of the different keying methods are shown, including the percent overlap for the weighted and the unit weight keys and also including test-retest reliability data for two keying methods, the current key and the best unit weight key.

The weighted keys were clearly superior to the unit weight keys, showing a lower overlap in all but two of the comparisons (with a few ties). But the difference was not great, never more than nine, usually only three to four percentage points.

These results can not be neatly categorized as to whether they are validation or cross-validation results. The eight occupational groups varied in the percentage of their composition that was the original validation group and the percentage that was used later for norming. If these were straight validation results, the difference in favor of the weighted system would be even less impressive as the validity of complex weighting systems shrinks more during cross-validation than that of unit weighting systems.

In either instance, the differences between the two scoring systems were not large enough to announce that either system is better. The advantage of simplicity in the unit weighting system overrides the slightly higher validity of the weighted system.

TABLE 2
COMPARISON OF OVERLAP RESULTS USING
400- VERSUS 298-ITEM POOL

Occupation	Percent of overlap between occupational group and men in general			
	Current key		Best unit weight key	
	400 Items	298 Items	400 Items	298 Items
Accountant	36	39	35	36
Author	32	33	30	31
Chemist	31	32	32	32
Engineer	40	40	45	44
Life insurance salesman	39	37	35	35
Production manager	46	46	48	49
Public administrator	37	36	37	37
Social science teacher	27	27	26	28

TABLE 3
RESULTS OF EXPERIMENTAL SCORING

Occupation	N	Current key ^a			Experimental keys ^b										
		Num- ber of items	Percent overlap		r _{tr}	± 4,3,2,1 items				± 4,3,2 items				± 4,3 items	
			Weighted	Unit weight		Num- ber of items	Percent overlap ^c	r _{tr} ^d	Num- ber of items	Percent overlap	Unit weight	r _{tr} ^d	Num- ber of items	Percent overlap	Unit weight
Accountant	345	191	36	43	88	156	39	46	23	34	36	84	10	37	35
Author	247	333	32	37	92	259	33	37	126	30	32	91	54	30	31
Chemist	297	286	31	38	93	226	32	39	54	28	32	91	23	30	32
Engineer	614	263	40	46	95	211	40	44	57	43	44	93	15	45	46
Life insurance salesman	310	260	39	47	91	211	37	46	48	28	35	90	10	26	26
Production manager	216	203	46	48	88	168	46	49	87	48	19	47	7	54	52
Public administrator	549	244	37	41	86	194	36	40	35	36	39	87	4	48	48
Social science teacher	217	297	27	35	89	229	27	34	44	23	28	87	21	25	28

Occupation	N	Current key ^a			Experimental keys										
		Num- ber of items	Percent overlap		r _{tr}	+ 4,3,2,1 items				+ 4,3,2 items				+ 4,3 items	
			Weighted	Unit weight ₂		Num- ber of items	Percent overlap	r _{tr}	Num- ber of items	Percent overlap	Unit weight	Num- ber of items	Percent overlap	Unit weight	
Accountant	345	191	36	43	88	131	39	48	20	36	41	7	38	40	
Author	247	333	32	37	92	239	38	41	111	36	40	48	36	38	
Chemist	297	286	31	38	93	202	36	41	39	31	36	16	30	31	
Engineer	614	263	40	46	95	192	43	47	48	44	47	13	47	47	
Life insurance salesman	310	260	39	47	91	185	40	48	42	31	38	10	28	29	
Production manager	216	203	46	48	88	130	52	58	15	45	49	4	47	49	
Public administrator	549	244	37	41	86	186	32	37	29	41	44	4	52	52	
Social science teacher	217	297	27	35	89	207	28	36	36	29	35	16	28	31	

Note.—Bold face numbers are best unit weight keys.
^a 400 items.
^b Using 298 unchanged items.
^c Occupational group versus Strong's MIG.
^d Based on unit weights.

TABLE 4
SCALE INTERCORRELATIONS USING CURRENT, BEST WEIGHTED,
AND BEST UNIT WEIGHT KEYS

Occupation	Accountant	Author	Chemist	Engineer	Life in- surance sales- man	Produc- tion manager	Public admin- istrator	Social science teacher
Accountant		−67 ^a	−05	−01	−12	44	33	28
		−46 ^b	03	01	09	35	28	11
		−45 ^c	00	07	07	35	31	14
Author			12	02	12	−47	−16	−27
			04	−05	−23	−54	01	01
			13	−09	−14	−46	−04	02
Chemist				72	−68	33	15	−26
				76	−65	39	24	06
				86	−76	40	22	−05
Engineer					−68	57	−09	−56
					−68	57	−01	−35
					−70	61	17	−18
Life insurance salesman						−29	−11	17
						−19	07	20
						−32	02	28
Production manager							15	−30
							23	−13
							19	−19
Public administrator								61
								59
								52
Social science teacher								
Key		Average intercorrelation (ignoring sign)		Number of times lowest of three				
Current		.31		8½				
Best weighted		.26		11				
Best unit weight		.28		9½				

^a Current key.
^b Best weighted key.
^c Best unit weight key.

The data in Table 3 also allows some comparison between keys of various lengths. Keys with roughly 40–80 items were better than longer ones. This seems to be true, no matter what method of item selection or weighting is used. It is not clear whether this is because there is some optimal upper limit on the number of items that a key should contain—Clark (1961) in an earlier report has taken this stand (p. 26)—or

whether there are usually only about this number of good items in a given occupational area.

The next step was to compare the reliability of the best unit weight keys with that of the current keys. To do this, 102 individuals took the SVIB twice with a 30-day interval.³ The test-retest reliabilities (r_{tr})

³ The 102 individuals were composed of two groups: 86 officers and men of the 360th Psycholog-

are reported in Table 3. Again, it appeared that the current weighted keys were slightly better than the best unit weight keys; but, again, the differences were small to the point of insignificance. In seven of the eight instances the weighted keys were slightly more reliable, but never more than four correlational points and usually by only one or two correlational points.

After the above validity and reliability data were available, three different keys for each occupational scale were selected for further study of the effect of the different scoring methods on the scale intercorrelations. The three were: the current key, the best weighted key, and the best unit weight key.

Other things being equal, the keying method which will give the lowest scale intercorrelations is preferred and the data were analyzed with that in mind. Reducing the number of items likely will have some effect on the relationship between the scales, simply by reducing the number of common items between scales. Table 4 contains the intercorrelations for the three types of keys for each.

There was not much difference in the size of the intercorrelations from one method to another. The best weighted key has a slight edge over the other methods, but not enough to be concerned about. These scale intercorrelations are not a crucial issue in deciding which keying method should be used.

One other kind of correlational data was calculated, the intercorrelations between the various experimental keys within one occupational group. Specifically, the author's scale was used within the criterion group of authors. From that intercorrelation table (not presented here) three conclusions were drawn:

1. The average intercorrelation between keying methods was extremely high, somewhere in the high .90s. The only low correlations, that is, between .80 and .90, were between scales with a low number of items, roughly 20 or less.

2. The lowest correlation between keys using weighted items and keys using the same items scored with unit weights was .98. Most of them were .99+.

3. The correlations between keys using the original 400 items and the same keys using only the 298 unchanged items were equally high, the lowest being .98, most of them .99+.

In general, these correlations were also of little help in deciding which keying method to use.

CONCLUSIONS

From all of the foregoing, the authors conclude:

1. Scoring the SVIB using only the 298 revised items will be as effective as using the original 400 items. Nothing will be lost by dropping the 102 obsolete or invalid items.

2. The weighting system of the SVIB should be revised to use only plus and minus 1 weights. While the data indicate that the weighting system was slightly superior to the unit weighting system, the differences were very small. The psychometric advantages of these differences are outweighed by the practical disadvantages created by the more complex weighting system.⁴

3. The items for the new scales should be selected from those items on the scales currently carrying the highest weights, and each new scale should have a minimum of 40

⁴ It seems desirable to note Strong's mild dissension. In a letter in 1962 he said,

I just don't seem to be able to come out flat-footedly regarding the use of weighted and unweighted items. For years I have favored the weighted items and still prefer them because they are more valid, but I have listened to a good many people who want the unweighted items because of the greater convenience. . . . I think I am going to pass the buck and let you people at Minneapolis decide which weighting is to be used.

ical Warfare Company, Ft. Snelling, Minnesota who participated through the cooperation of their Commanding Officer, Major Walter Angrist, and 16 acquaintances of the second author. On their stability from test to retest, the latter group—who knew what the authors were about—proved to be more reliable. However, they were left in the analysis as any bias introduced should equally affect all methods of keying. Elmer Hanks of TestScor provided the scoring of these test-retest blanks, a service the authors gratefully acknowledge.

items. Specifically, for a given scale, if there are more than 40 items with ± 4 weights, use those. If there are less than 40, add the ± 3 weighted items, and so fourth, progressing until the scale has at least the minimum of 40.

As the results in Table 3 indicate, usually this will mean using the items weighted ± 4 , ± 3 , and ± 2 . In some cases, however, it will mean using only ± 4 and ± 3 ; in others, all of the items ± 4 through ± 1 will be used.

REFERENCES

- CLARK, K. E. *Vocational interests of nonprofessional men*. Minneapolis: Univer. Minnesota Press, 1961.
- FRUCHTER, B. Preferred selection, scoring, and weighting procedure for background and preference items. Paper read at the American Educational Research Association, Chicago, February 1963.
- KUDER, G. F. A rationale for evaluating interests. *Educ. psychol. Measmt*, 1963, **23**, 3-12.
- STRONG, E. K., JR. Good and poor interest items. *J. appl. Psychol.*, 1962, **46**, 269-275.

(Received September 1, 1963)

RELATION BETWEEN RADAR DETECTION AND THE OBSERVER'S CONCEPT OF A TARGET

R. D. BALDWIN, A. D. WRIGHT, AND D. J. LEHR

Human Resources Research Office, Fort Bliss, Texas

An experiment tested the hypothesis that target detectability on a PPI radar display depends on O's knowledge of the attributes defining a target. Equal numbers of O's were given either a brightness, a form, or a combined brightness-form set during training. A 4th group was given only demonstration training. The criterion test involved detection of 2 target sizes in 2 levels of visual noise for 3 target speeds. Analysis of variance revealed an interaction between set and noise level, confirming the hypothesis for the high noise level only.

Under optimum radar operating conditions, the target signal, or "pip," on a plan position indicator is large and well contrasted with the background video noise level. Under nonoptimum conditions of strong radio frequency interference and small targets, the radar observer has a difficult discrimination task. The target pip is small (1×2 millimeters for some radars), the form is not well bounded, the pip will be painted intermittently, and the brightness contrast will be low. Under such impoverished stimulus conditions, Erdmann and Myers (1958) have suggested that target detectability will be dependent more on form discrimination than brightness discrimination. This hypothesis can be extended to the prediction that under impoverished conditions target detectability is dependent upon the observer's knowledge of the number of dimensions or attributes defining a target. An experiment was conducted to test this hypothesis that detection performance is directly related to the complexity of the observer's concept of a target.

METHOD

Training

Forty naive subjects were individually instructed by one of four training procedures to establish different sets concerning the defining attribute(s) of a radar target. Two attributes were used: target form and the brightness contrast. Two groups were given either a brightness or a form set, a third group was set for a combination of brightness and form, and a fourth group was given demonstration training only. For the latter group, no target attributes were mentioned during training. The instructor merely pointed

to the pip and called it "the target." Fifteen practice trials were given each observer. Knowledge of results combined with repetition of the relevant set was given for each practice trial.

Criterion Test

The criterion test consisted of 30 trials involving equal presentation frequencies of large and small targets (1-square meter and 100-square meter radar cross section) at three target air speeds (500, 700, and 900 knots). Equal numbers of observers from each training group were tested under either high or low levels of background video noise. Measurement of the dimensions of the target pip was not possible because of the dynamic nature of the signal which varied systematically with the range of the target and the electronics countermeasures (ECM) level. The root mean square (RMS) noise amplitude for the high level noise condition was approximately 1.35 volts while RMS noise amplitude for the low level noise was approximately 1.05 volts. Target amplitude varied inversely as the fourth power of target range and was approximately 1.75 volts and .9 volt for the small targets at 50,000 and 100,000 yards range and 2.5 volts and 1.5 volts for the large targets at the respective ranges.

Apparatus

Simulated targets and electronic interference (video noise) were displayed on a 10-inch P-7 phosphor Plan Position Indicator (PPI) installed in a radar control system. Targets were programed at a single heading on radial tracks. The distance of the target from the center of the PPI at initial appearance depended upon the particular combination of target size and noise level used for each trial. The observers' scores were adjusted for this variation in range of initial appearance by a correction constant based on the average range of the targets when detected by two instructors. These calibration trials were conducted for each size target immediately before each observer was trained.

RESULTS

The average corrected scores (distance from point of initial appearance in yards) obtained by each training group for each combination of target size, speed, and noise level are presented in Table 1.

The scores were analyzed employing an extension of the Lindquist (1953) Type VI mixed analysis of variance. The summary is presented in Table 2.

The between-subjects analysis yielded a reliable interaction between noise level and training set. For the high noise level, the smallest error scores were obtained by the subjects given the combined brightness-form set. Under the low noise condition, the subjects given the form set had the best scores and the brightness-form group had the largest errors. Supplementary variance analysis indicated that the variation among the means for the low noise condition was not reliable ($F = 2.27$; $df = 3/32$), whereas significant differences did exist for the high noise condition ($p < .025$).

TABLE 1
MEAN ADJUSTED DETECTION SCORES^a

Sets	Average: all sizes- speeds	Target size	Target speed (knots)		
			500	700	900
High ECM level					
Form	3210	Large	180	1,520	3,440
		Small	4,400	6,960	2,760
Brightness	4807	Large	3,320	3,720	7,040
		Small	3,400	6,000	5,360
Brightness- form	2100	Large	4,240	4,320	3,360
		Small	−1,720	−520	2,920
Control	8667	Large	4,240	8,120	10,320
		Small	6,640	10,760	11,920
Low ECM level					
Form	1013	Large	920	3,400	−80
		Small	600	−480	1,720
Brightness	3833	Large	2,000	5,440	3,280
		Small	3,080	3,200	6,000
Brightness- form	6287	Large	5,400	5,040	6,920
		Small	6,440	5,280	8,640
Control	4173	Large	3,880	3,480	5,440
		Small	3,720	3,120	5,400

^a Yards after burnthrough. A negative score indicates that the target was seen before the instructors' burnthrough point. A low score indicates superior performance.

TABLE 2
ANALYSIS OF VARIANCE

Source	df	MS	F
Between subjects			
Training (T)	3	4,361	2.82
Noise level (N)	1	1,316	.85
T × N	3	5,155	3.33*
Error (between)	32	1,548	
Within subjects			
Target speed (V)	2	2,012	13.16****
Target size (S)	1	66	.14
V × S	2	69	.50
V × T	6	394	2.58*
V × N	2	318	2.08
S × T	3	445	.91
S × N	1	9	.02
V × S × T	6	113	.82
V × S × N	2	775	5.62***
V × T × N	6	192	1.26
S × T × N	3	1,215	2.48
V × S × T × N	6	383	2.78**
Error 1 (V terms)	64	153	—
Error 2 (S terms)	32	489	—
Error 3 (VS terms)	64	138	—

* $p < .05$.
** $p < .025$.
*** $p < .01$.
**** $p < .001$.

The within-subjects analysis revealed a reliable four-fold interaction between the training, target size, speed, and noise level factors. Additional analyses were made for each target size-noise level combination with the following results:

1. There were no reliable differences among the mean detection scores of the four training groups for large targets in either high or low noise or for small targets in low noise.
2. Reliable differences in detection scores occurred for small targets in high noise with the brightness-form group obtaining the best performance.
3. Considerable variability existed in the effect that target speed had on detections. There was a general tendency for error magnitude to increase with target speed. However, a considerable and unaccountable number of reversals in this trend occurred for specific target size, noise level, and training combinations.
4. Contrary to usual findings, the analysis did not indicate that the noise level had a

reliable general effect on performance. This result was due to the method of correcting the scores which eliminated the constant effects of noise level.

DISCUSSION

The finding that the group given the combined instructions had the earliest detections under the high ECM condition confirms the hypothesis that under impoverished stimulus conditions radar detection is directly related to the dimensional complexity of the observer's concept of a target. These results also support the Erdmann and Myers hypothesis that the target's form is a critical attribute influencing detection under visual noise conditions.

The paradoxical effect of the combined set for the two noise levels presents a problem for interpretation. Target detection can be regarded as a decision task. As such, the detection score, yards distance from initial appearance, can be treated as a measure of decision latency. For each scan of the radar antenna, the observer must decide if a stimulus component is present that possesses a target attribute. In terms of single stimulus attributes, the results of this experiment for

both low and high ECM suggest that this decision process requires less time if the observer is set for object form rather than brightness contrast. In addition, from Table 2, Column 2, it is apparent that the average score obtained by the combined set group under low ECM can be approximated by summing the mean scores of the single attribute groups. This observation suggests that for the low noise condition, observers given the combined training made two independent decisions for each scan of the antenna: one for form and a second for brightness. However, this line of reasoning does not hold for the high noise condition, since the average latency for the combined set condition was less than that for either single attribute set.

REFERENCES

- ERDMANN, R. L., & MYERS, R. D. The effect of number of signal pulses upon signal detectability with PPI scopes. In, *Illumination and visibility of radar and sonar displays*. (Publ. No. 595) Washington, D. C.: National Academy of Sciences-National Research Council, 1958.
- LINDQUIST, E. F. *Design and analysis of experiments in psychology and education*. New York: Houghton Mifflin, 1953. Pp. 292-297.

(Received February 28, 1963)

APPLICATION OF A METHOD OF EVALUATING TRAINING¹

JOHN A. COX

Human Resources Research Office, Fort Bliss, Texas

Data were processed with Ward Edwards' formulation of value of training which includes estimates of proficiency level attained, worth of a trained man in dollars, and training costs in dollars. Difficulties which were encountered and techniques of overcoming them are reported. Results of the evaluation, which appear to be realistic, are reported.

A method of evaluating training methods has been proposed by Edwards (1956). This method incorporates proficiency gain, worth of training, and cost estimates into a single figure. Such simplicity holds marked appeal, but the method has not been used, as far as can be determined.

An opportunity occurred to gather data from an ongoing training program and to process these data with Edwards' technique. This article reports the results.

EVALUATION TECHNIQUE

Edwards points out that there are three functions which need to be considered when evaluating a training program: proficiency level of a trained man, cost of a training program which produces this proficiency level, and worth of a man trained to this proficiency level. When two training programs are compared, there are two levels of proficiency and two costs to be compared. Edwards formulates this comparison process as follows:

$$G = n \sum (p_x - p_c)R - n(C_x - C_c)$$

where:

G = gain of one training program over another in terms of dollars

n = number of persons studied

p_x = proportion of students in X Group above a given level of proficiency

p_c = proportion of students in C Group above a given level of proficiency

R = worth of a trained man

C_x = cost of training program X

C_c = cost of training program C

¹ Data for this study were collected through the cooperation of United States Army Signal Fire Distribution Systems Training Activity and instructor personnel from Hughes Aircraft Company, Ground Systems Group, Fort Bliss, Texas.

To the above conceptualization was added the concept of sampling error. By computing the standard error of the difference between the two proportions ($p_x - p_c$), two estimates of gain can be computed. A multiple of the standard error figure must be added to the difference figure for an estimate of maximum gain and subtracted from the difference figure for an estimate of minimum gain.

TRAINING SITUATION

In a 21-week maintenance course for electronic equipment there was a change to be made. This change affected 2 hours of class time by adding the use of a small training device to the instructional procedure and reducing the amount of lecture and blackboard illustration. This situation provided the opportunity to apply Edwards' technique to data collected from men trained with the lecture and blackboard method and the training device method.

DATA COLLECTION

To get proficiency scores an achievement test was constructed. The test covered the 2 hours of instruction only. A class of 20 students became available for testing about 3 weeks after the men had completed the 2-hour instruction period. This class had received lecture and blackboard instruction. This class was administered the criterion test and became Group C for this study. The next class was trained with the training device. This class of 24 men also was tested with the criterion test 3 weeks after completing the 2-hour instruction period. This class became Group X for the study. Each of these men qualified for entry into electronics training in the United States Army and had graduated from a basic electronics course.

Cost information was furnished by the school administration. Costs for this course were \$32.05 per student per day. Each man recorded his rank and date of enlistment on his test form as identification data. Information as to pay and re-enlistment rates was obtained from the post recruiting office.

DATA PROCESSING

For each group, the mean proficiency score was computed. The proportion of scores in each group which fell above the mean of Group C was computed. These figures are the p_x and p_c of Edwards' formula. The difference between these proportions and the standard error of that difference were computed. To approximate the 95% confidence interval, this standard error figure was doubled and added to the difference. Then the standard error was doubled and subtracted from the difference. All the above figures are presented in Table 1.

Only rough estimates of the worth of a trained man were attempted. The rationale for this procedure was as follows. A man is worth more than you pay him. If he wasn't, you wouldn't hire him. We pay soldiers who are "less than E-6 rank" less than \$15.00 per day. So we will estimate that each of these men, on the average, is worth at least \$15.00 per day. We pay soldiers who are "E-6 rank and over" less than \$25.00 per day. So we will estimate that each man whose rank is E-6 or higher is worth at least \$25.00 per day. Each of these men, after training, will go to a job assignment where he will use the training he received and

TABLE 1
SUMMARY OF CALCULATIONS

	Group C	Group X
<i>N</i>	20	24
<i>M</i>	30.40	35.88
<i>p</i> above <i>M_c</i>	.500	.667
E-6 and up	5	3
<i>p_x - p_c = .167</i>		
<i>SE_{diff.} = .148</i>		
<i>p_x - p_c + 2SE = .757</i>		
<i>p_x - p_c - 2SE = -.423</i>		

will remain on such assignment for at least 2 years. If he is in his first enlistment, the higher probability is that he will not re-enlist and about 2 years after his training is completed, the Army will no longer receive any benefit from the investment in training. If the man is a re-enlistee, he will probably re-enlist for his own vacancy. Later, he will be promoted or transferred so that about 4 years after his training is completed, he is no longer on the specific job for which he was trained. While the Army may continue to receive some benefit from the training investment in this man, the value after 4 years is markedly decreased. The above rationale tends to agree with reality but makes many assumptions. And even the "real" part will show many differences, soldier by soldier. But this is the rationale used in this study.

Figures from the above discussion, along with appropriate ones for days and years, were entered into the following formula:

$$R = H_s \left[\frac{[4(365)][RE-6(\$25.00) + RE-5(\$15.00)] + [2(365)][EE-6(\$25.00) + EE-5(\$15.00)]}{\frac{N}{H_c}} \right]$$

where:

EE-6 = number of men E-6 and above in first enlistment

RE-6 = number of men E-6 and above re-enlisted

EE-5 = number of men E-5 and below in first enlistment

RE-5 = number of men E-5 and below re-enlisted

N = total number of men in this study

H_c = total number of hours in this course

H_s = total number of hours in this study

For this study the figures in the formula were:

EE-6 = 0

RE-6 = 8

EE-5 = 27

RE-5 = 9

N = 44

H_c = 840

H_s = 2

when the equation was solved, $R = \$42.46$ —the estimated worth to the Army of a man trained for this 2 hours.

The school furnished a basic cost estimate for this course of \$32.05 per student per day. Using this estimate, 2 hours of training costs \$8.01. The training device cost was prorated for this class of men and added to the above figure to give a cost estimate for Group X of \$9.15 per man.

Two gain estimates could now be computed, one minimum and one maximum. The figures for minimum estimate of difference in proficiency (see Table 1), worth of a man, and cost were entered in the Edwards formula, and the equation was solved. The minimum expected gain which resulted from this computation was $-\$840.43$. This means that if the training device was used regularly in this program, the Army could reasonably expect not to lose more than \$840.43 per 44 men trained. Changing to maximum expected gain figures, the equation was solved again, producing an estimate of \$1364.00. This means that if the training device was used in the course regularly, the Army could reasonably hope to gain a net of no more than \$1364.00 per 44 men trained.

DISCUSSION

The most troublesome part of the procedure was the worth-of-a-man estimate. The computations were not difficult, but arriving at a procedure which would produce a figure that had some degree of reality was elusive. An early procedure was based on all 21 weeks of training and assumed that a re-enlistee would spend 20 years in the service and that the Army would receive full benefit from this training for that time. The results, when processed through the formula, estimated a potential loss of almost \$1,000,000, and a potential gain of as much as \$1,500,000. This is a little too much to expect from a 2-hour training episode, either loss or gain. The loss was more questionable because it assumed that the lower proficiency man never gained proficiency after leaving training. Actually there is research evidence that Army maintenance personnel continue to learn on the job for at least the first year after training. For the trouble-shooting part of a job, they continue to gain in proficiency for even longer periods (Hitchcock, Mager, & Whipple, 1958).

The worth-of-a-man estimate was revised so that the basis was for 2 hours of training and for a 2-year or 4-year posttraining period.

Training cost estimates underwent a similar evolution. However, the process was less involved and the errors were more obvious. Probably the most effective technique of solving these problems is to think of them in units of "one man for one training period." When this approach was taken, estimates that looked somewhat realistic were attained.

The problem of "negative gain" still remains in the system proposed here. In the situation reported, proficiency gain due to application of the training device was so small that on a statistical, or probability, basis one would expect that in some future classes the observed gain would disappear and even be replaced by a loss. In other words, there is a probability that some future classes taught with the training device would show a lower proficiency level than would some classes not taught with the training device. The statistical model assumed that, after training, men learn nothing. That assumption is largely false. Men do learn on the job—at least many of them do. At present, there is no obvious systematic solution to this problem of negative gain. If the statistical model is used, negative gain is likely to occur in any similar experiment or evaluation. If the statistical model is not used (but instead the real figures from the sample which happened to be available for study are used) then any gain, no matter how small and unlikely to recur, would be endorsed by the study and would recommend that a training device be used. The lesser of these evils seems to be the statistical model.

As proposed by Edwards and tried and reported here, the process gives results in terms of the number of men studied. In the present case, this was 44 men. In another study it might be 168 or any other number. A more useful procedure would seem to be to compute gain estimates for a standard number, say 1 or 100. This would simplify computation and interpretation of results. The formula can be modified to:

$$G = (p_x - p_c)(R) - (C_x - C_c)$$

to show the expected gain for one trained man. The data collected in this study were processed

with the above formula. The estimates per one man trained were: minimum estimated gain = -\$19.10, maximum estimated gain = \$31.00. These are net gains expected from teaching the 2 hours of material with the training device. Each man so taught would produce this much net gain. And the gain would occur over a 2- to 4-year period, after which no more gain to the Army is expected.

The question remains, "Is Edwards' technique, as used here, useful and appropriate?" The author's opinion is that the answer is "Yes"—granted that some difficulty was experienced in coming to grips with the "estimate of the worth of a trained man." The assumptions and rationale used here, and stated by Edwards, underlie any similar evaluation of training, even though the evaluator may not acknowledge them overtly. Since this situation exists, overt acknowledgment and a deliberate attempt to incorporate an estimate of them into an evaluation process seem proper. There is some frustration experienced when such rough estimates are required. However, there is some gratification experienced when a com-

parison can be made in terms of dollars and cents. This procedure, or some modification of it, should be especially useful to military and industrial training personnel. These persons are quite frequently required to justify money expenditures for their training programs. The process studied here will promote objectivity in evaluating and justifying these training programs. Programs of public education, as opposed to those of training, will likely find this process much less useful. "Worth of a trained man" and "cost" estimates should be much more elusive to public education personnel.

REFERENCES

- EDWARDS, W. The use of statistical decision functions in making practical decisions. In, *Symposium on air force human engineering, personnel, and training research*. Washington, D. C.: National Academy of Sciences-National Research Council, 1956. Pp. 115-124.
- HITCHCOCK, L., JR., MAGER, R. F., & WHIPPLE, J. E. Development and evaluation of an experimental program of instruction for training fire control technicians. *USA Air Def. Hum. Res. Unit tech. Rep.*, 1958 (May), No. 46.

(Received March 7, 1963)

SEX DIFFERENCES IN JOB SATISFACTION¹

CHARLES L. HULIN

University of Illinois

AND

PATRICIA CAIN SMITH

Cornell University

Measures of 5 separate aspects of job satisfaction gathered from 295 male workers and 163 female workers drawn from 4 different plants were analyzed with respect to the mean job satisfaction for the male and female workers. T² analyses indicated that in 3 plants the female workers were significantly less satisfied than their male counterparts ($p < .05$) while in the 4th plant there was no significant difference. A test of concordance on the relative size of the differences indicated that the ordering of the differences in satisfaction level was somewhat consistent across the 4 samples ($p < .01$).

What are the antecedent conditions which lead to high or low job satisfaction? Various writers have investigated the contribution of job level to satisfaction (Ash, 1954; Centers, 1948; Hulin, 1963; Katz, 1949), the contribution of wage level (Hulin, 1963; Inlow, 1951; Miller, 1940, 1941; Survey Research Center, 1950), the effects of type of leadership or leadership climate (Fleishman, 1955; Kahn & Katz, 1953; Pelz, 1952; Vroom & Mann, 1960), the age of the worker as it affects job satisfaction (Bernberg, 1954; Cain, 1942; Hinrichs, 1962; Hulin, 1963; Mann, 1953; Smith, 1955), and even the effects of the personality of the workers on their job satisfaction (Cain, 1942; Peterson & Stone, 1942; Smith, 1936; Smith, 1955; Weitz, 1952).

The net result of these studies seems to be that higher job levels and higher wages generally contribute to higher job satisfaction; the type of leadership has certain effects on job satisfaction, but these effects are modified greatly by situational factors; age and tenure seem to be positively related to job satisfaction (for a dissenting opinion on this latter conclusion see Herzberg, Mausner, Peterson, & Cap-

well, 1957); and job satisfaction seems to be related to a general life adjustment-maladjustment factor.

The relation of the sex of the worker to job satisfaction is an additional topic which has received a great deal of attention. The findings of the investigations on sex differences in job satisfaction, however, are somewhat contradictory and permit no neat cogent statement of the relationship between sex and job satisfaction.

For example, Bengé (1944) and Stockford and Kunze (1950) concluded that women are *more* satisfied than men, while Cole (1940) reported women to be *less* satisfied than men, and Blood, Harwood, and Vernon (1942), in a study of the psychological problems of adjusting to the rather severe wartime working conditions which existed in Great Britain during World War II, reported that women exhibited more serious psychological maladjustment than did their male counterparts.

In the area of satisfaction with teaching, Chase (1951) reported women teachers to be more satisfied than men, while Peck (1936) concluded that women were more poorly adjusted than men teachers.

Not only do these results contradict each other to a considerable extent, but also some of them contradict the findings, reported above, concerning wages and job level and their effects on job satisfaction. Based on these findings, women should be less satisfied than men, since they are usually placed on lower level jobs, which have a lower pay rate, and which usually offer few promotional opportunities. Thus, barring the possible effects of adaptation level, women should be less satisfied on their jobs.

¹ This study is part of the Cornell University Studies of Retirement Policies financed by a grant from the Ford Foundation and is based on a portion of a doctoral dissertation by the senior author submitted to the graduate school of Cornell University in partial fulfillment of the requirements for the degree of Doctor of Philosophy. The dissertation was directed by the second author. Both authors participated in collection of the data.

We wish to express our gratitude to the cooperating companies who made their records available and contributed the time of their personnel to make these studies possible, and to the interviewers who contributed their time to the validation studies.

In addition to the factors of wages and job level, there is the issue of societal norms concerning appropriate roles for men and women. When males are employed in industry they are filling the role that society has come to expect of them. Women in industry (in spite of their increasing numbers) are in a relatively alien role. In addition, if they are married and working full time they may be faced with a certain amount of role conflict, which also may affect their job satisfaction. (See Mead, 1949, for a discussion of some of these problems of dual roles which female workers must face when they hold a full-time job.)

It should be noted that there may be some methodological problems connected with many of the studies of sex differences. In general the questionnaires which have been used to measure job satisfaction seem to have been designed for, and "validated" on, male workers. With the known differences in the motivation of male and female workers (Jurgensen, 1949) this practice may introduce a bias of unknown but substantial magnitude. An additional problem is the fact that these investigations all used different and not necessarily equivalent measures of job satisfaction (see Smith & Kendall, 1963). Thus, we do not know if the apparently contradictory results are due to differences in the measures used or to true differences in satisfaction levels between men and women.

The present investigation reports an analysis of data of workers drawn from four different plants representing three different companies. The same questionnaire was administered to all of these workers. Thus, one source of the apparent contradiction may be eliminated. Secondly, the measures of job satisfaction used were validated and item analyzed using both male and female subjects. (This questionnaire will be described in more detail below.)

METHOD

Sample

The workers included in this analysis were given a guarantee of complete anonymity in all cases. It was stressed that their results would never be made known to the company officials.

They were drawn by means of a systematic sample from the employment rolls of the companies involved. (a) Company I, a large electronics firm located in a large metropolitan area in New England. Two units of this company, Plants A and B, were studied with each

plant employing between 1,000 and 1,500 workers. Sampling was stratified by age, with no further restrictions. (b) Company II, a manufacturer of cardboard products, a medium-sized (300 to 400 employees) family-owned plant located in a small city in New England. Sampling was alphabetical. (c) Company III, a brass foundry, a medium-sized company (300 to 400 employees) located in a large city in the Midwest. Sampling was by clock number.

The sampling procedure used yielded a sample of 99 men and 35 women from Plant A, Company I; and 86 men and 40 women from Plant B, Company I. Company II gave a sample of 50 men and 43 women. Company III gave a sample of 60 men and 25 women. The analysis to be presented in this paper is based on a total sample of 295 male workers and 163 female workers.

Measures of Job Satisfaction

The device used to measure satisfaction in this study had been developed as part of a nationwide survey of retirement satisfaction. It had been subjected to an intensive validation program. The results of these validation studies were presented by Smith (1961, 1963), Hulin (1961), Hulin, Smith, Kendall, and Locke (1963), Macaulay (1961), Macaulay, Smith, Locke, Kendall, and Hulin (1963), Kendall (1961), Kendall, Smith, Hulin, and Locke (1963), Locke (1961), and Locke, Smith, and Hulin (1963). The brief discussion presented here will not attempt to present the complete details of these studies but will give only a short summary of these papers.

Basically, the Job Descriptive Index (JDI) is an adjective check list on which each worker is asked to describe several aspects of his job by means of a "yes," "?," or "no" response to each of the adjectives. The aspects of the job which the workers describe are their work, their pay, their opportunities for promotion, their supervision, and the people with whom they work. These areas or aspects of the job were chosen so as to be consistent with the findings of the factor analytic studies which have been done on the dimensions of job satisfaction (Ash, 1954; Austin, 1958; Baehr, 1954, 1956; Baehr & Renck, 1958; Gordon, 1955; Harrison, 1960, 1961; Twery, Schmid, & Wrigley, 1958; Wherry, 1954, 1958).

Each response to each adjective in the final form of the JDI has been item analyzed against total scale score to determine the proper scoring direction. Items were retained which discriminated significantly for *both* male and female workers separately.

Before this final scoring key was developed, however, the investigators developed and used several other methods of scoring the JDI. In the early studies, each worker's JDI was scored in four different ways. Each of the four different resulting scores was correlated with several different job satisfaction ratings made by both interviewers and the workers themselves. Different scoring methods were systematically eliminated from consideration on the basis of these "validity" estimates. The final method of scoring and selection of items were those which yielded the highest estimates of convergent and discriminant validity (Campbell & Fiske, 1959)

TABLE 1
VECTORS OF MEAN DIFFERENCES

Company	Work	Area of job satisfaction				<i>p</i>
		Pay	Promotions	Supervision	People	
I, Plant A	.88	1.28	8.64	5.08	5.10	<.05
I, Plant B	-1.88	-3.28	2.92	.14	-1.18	<i>ns</i>
II	2.80	-.48	4.91	1.10	6.37	<.05
III	.00	-2.00	6.72	2.37	1.86	<.05

across several different samples for both male and female subjects.

In addition to the validity estimates, which averaged .50 to .70 across several samples, the JDI also has several other desirable characteristics.

1. The scores on it are unaffected by acquiescence or yes-saying and no-saying tendencies.

2. The resulting five scales, while not completely orthogonal, have the virtue of relatively low intercorrelations (.30 to .50) with each other.

3. Factor analyses of the data from several samples indicate that the workers are indeed capable of thinking along the lines of five separate aspects of job satisfaction. The factors extracted do seem to correspond to the five dimensions chosen by the investigators.

4. The five scales, while being quite short and easily administered, have adequate split-half reliabilities (.80 to .88 corrected by the Spearman-Brown formula).

RESULTS

Due to the multivariate nature of the measures of job satisfaction which were used in this study, the sex differences in job satisfaction could not be analyzed by the usual univariate test of significance. Instead Hotelling's T^2 analysis was used. This test is simply the multivariate analogue of the t test. The T^2 analysis, instead of using a single difference between means ($\bar{x}_i - \bar{x}_j$) and a variance estimate ($s_{x_i}^2 + s_{x_j}^2$), is based on a vector of mean differences ($\bar{x}_1 - \bar{x}_2, \bar{y}_1 - \bar{y}_2, \dots$) and the inverse of the variance-covariance matrix. This procedure has the advantage of providing an overall test of the significance of a vector of mean differences and of taking into account both the variance and the covariance of the variables as well as the magnitude of the difference. It thus avoids many of the problems connected with multiple comparisons. It has the disadvantage, however, of testing for the significance of the vector as a whole and providing no estimate of the contribution of the individual elements which make up the vector. For a complete

discussion of this statistical test see Hotelling (1931) or Anderson (1958, pp. 101-103).

Table 1 gives the vectors of differences which were obtained from the four plants studied. A positive element in the vectors indicates that the male workers were more satisfied with that aspect of their work than were the females.

It can be seen from this table that in three of the four plants the female workers were significantly less satisfied than the male workers. These findings would seem to support, to a limited extent, our hypothesis that female workers should be less satisfied than male workers. This finding is not completely general in our sample of plants since one plant indicated no differences between male and female workers.

It also should be noted that a test of concordance done on these data indicates that not only are the *levels* of satisfaction different between the two sexes, but the *relative ordering* of these differences remains reasonably constant across the four different samples ($w = .78, p < .01$). In three of the four plants the females were more dissatisfied (relative to the males) with their promotional opportunities than with any other factor of their jobs. This is about what one would expect considering their opportunities for promotions. In three of the four plants females were slightly more satisfied with their pay than males. It is also an objective fact that women get *less* pay than men. When their pay is considered relative to their job level, type of work, and economic needs, however, it may not be out of line with the men's pay.

DISCUSSION

The data we have presented, based on a sample of 295 male workers and 163 female

workers, seem to indicate that the female workers tend to be somewhat less satisfied with their jobs than their male counterparts. Further, the relative magnitude of the obtained differences was somewhat consistent across samples and was not out of line with what one might reasonably expect to obtain. There are two points, however, which merit further discussion. Firstly, we do not maintain that sex per se is the crucial factor which leads to either high or low satisfaction. It is, rather, the entire constellation of variables which consistently covary with sex; for example, pay, job level, promotion opportunities, societal norms, etc., that is likely causing the differences in job satisfaction. It is also likely that if these variables were held constant or if their effects were partialled out, the differences in job satisfaction would have disappeared. This study was intended primarily to establish the actual facts of the situation and not to offer an explanation. In the industrial setting these factors are *not* controlled or held constant and they do covary with sex. In each of the four samples the women were receiving less pay and were working on lower level jobs than the men, and in three of the samples the women were less satisfied than the men. The fact is that a large (and increasing) percentage of our work force is working under the handicap of relative dissatisfaction.

A second important point is that the difference in satisfaction did not hold up across all four samples, even though (as mentioned above) in all four samples the women were working on lower level jobs and were receiving less pay. This failure could perhaps be attributed to sampling fluctuations. On the other hand it is more likely that situational factors play a very important role, not only in the level of satisfaction of the workers, but also in the relative satisfaction of the male and the female workers. Thus, even though we can conclude now that women are generally less satisfied than men, additional precision concerning this conclusion could likely be gained by considering situational factors.

REFERENCES

- ANDERSON, T. W. *Introduction to multivariate statistical analysis*. New York: Wiley, 1958.
- ASH, P. The SRA employee inventory: A statistical analysis. *Personnel Psychol.*, 1954, 7, 337-364.
- ASTIN, A. W. Dimensions of work satisfaction in the occupational choices of college freshmen. *J. appl. Psychol.*, 1958, 42, 187-190.
- BAEHR, MELANY E. A factorial study of the SRA employee inventory. *Personnel Psychol.*, 1954, 7, 319-336.
- BAEHR, MELANY E. A reply to R. J. Wherry concerning "An orthogonal re-rotation of the Baehr and Ash studies of the SRA employee inventory." *Personnel Psychol.*, 1956, 9, 81-92.
- BAEHR, MELANY E., & RENCK, R. The definition and measurement of employee morale. *Admin. Sci. Quart.*, 1958, 3, 157-184.
- BENGE, E. J. How to learn what workers think of job and boss. *Factory Mgmt. Maint.*, 1944, 102(5), 101-104.
- BERNBERG, R. E. Socio-psychological factors in industrial morale: III. Relation of age to morale. *Personnel Psychol.*, 1954, 7, 395-399.
- BLOOD, W., HARWOOD, J., & VERNON, H. M. Discussion on effects of war-time industrial conditions on mental health. *Proc. Roy. Soc. Med.*, 1942, 35, 693-698.
- CAIN, PATRICIA A. Individual differences in susceptibility to monotony. Unpublished doctoral dissertation, Cornell University, 1942.
- CAMPBELL, D. T., & FISKE, D. W. Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychol. Bull.*, 1959, 56, 81-105.
- CENTERS, R. Motivational aspects of occupational stratification. *J. soc. Psychol.*, 1948, 28, 287-317.
- CHASE, F. S. Factors for satisfaction in teaching. *Phi Delta Kappan*, 1951, 33, 127-132.
- COLE, R. J. A survey of employee attitudes. *Publ. Opin. Quart.*, 1940, 4, 497-506.
- FLEISHMAN, E. A. Leadership climate, human relations training, and supervisory behavior. *Personnel Psychol.*, 1955, 6, 205-222.
- GORDON, O. J. A factor analysis of human needs and industrial morale. *Personnel Psychol.*, 1955, 8, 1-18.
- HARRISON, R. Sources of variation in managers' job attitudes. *Personnel Psychol.*, 1960, 13, 425-434.
- HARRISON, R. Cumulative communality cluster analysis of workers' job attitudes. *J. appl. Psychol.*, 1961, 45, 123-125.
- HERZBERG, F., MAUSNER, B., PETERSON, R. O., & CAPWELL, DORA F. *Job attitudes: Review of research and opinion*. Pittsburgh, Pa.: Psychological Service of Pittsburgh, 1957.
- HINRICHS, J. R. The impact of industrial organization on the attitudes of research chemists. Unpublished doctoral dissertation, Cornell University, 1962.
- HOTELLING, H. The generalization of student's ratio. *Ann. math. Statist.*, 1931, 2, 360-378.
- HULIN, C. L. Cornell studies in methods of measuring job satisfaction: I-A systematic approach to the measurement of job satisfaction. Paper read at American Psychological Association, New York, 1961.
- HULIN, C. L. A linear model of job satisfaction. Unpublished doctoral dissertation, Cornell University, 1963.
- HULIN, C. L., SMITH, PATRICIA C., KENDALL, L. M., & LOCKE, E. A. Cornell studies of job satisfaction:

- II. Model and method of measuring job satisfaction. Ithaca: Cornell University, 1963. (Mimeo)
- INLOW, G. M. Job satisfaction of liberal arts graduates. *J. appl. Psychol.*, 1951, **35**, 175-181.
- JURGENSEN, C. E. What job applicants look for in a company. *Personnel*, 1949, **25**, 352-355.
- KAHN, R., & KATZ, D. Leadership practices in relation to productivity and morale. In D. Cartwright & A. Zander (Eds.), *Group dynamics*. Evanston, Ill.: Row, 1953.
- KATZ, D. Morale and motivation in industry. In W. Dennis (Ed.), *Current trends in industrial psychology*. Pittsburgh: Univer. Pittsburgh Press, 1949.
- KENDALL, L. M. Cornell studies in methods of measuring job satisfaction: III. The relative validity of different methods of measurement for predicting criteria of satisfaction. Paper read at American Psychological Association, New York, 1961.
- KENDALL, L. M., SMITH, PATRICIA C., HULIN, C. L., & LOCKE, E. A. Cornell studies of job satisfaction: IV. The relative validity of the job descriptive index and other methods of measurement of job satisfaction. Ithaca: Cornell University, 1963. (Mimeo)
- LOCKE, E. A. Cornell studies in methods of measuring job satisfaction: II. Importance and satisfaction in several job areas. Paper read at American Psychological Association, New York, 1961.
- LOCKE, E. A., SMITH, PATRICIA C., & HULIN, C. L. Cornell studies of job satisfaction: V. Scale characteristics of the job descriptive index. Ithaca: Cornell University, 1963. (Mimeo)
- MACAULAY, D. ANNE. Cornell studies in methods of measuring job satisfaction: Development of criteria of job satisfaction. Paper read at American Psychological Association, New York, 1961.
- MACAULAY, D. ANNE, SMITH, PATRICIA C., LOCKE, E. A., KENDALL, L. M., & HULIN, C. L. Cornell studies in job satisfaction: III. Convergent and discriminant validity for measures of job satisfaction by rating scales. Ithaca: Cornell University, 1963. (Mimeo)
- MANN, F. C. A study of work satisfaction as a function of the discrepancy between inferred aspirations and achievement. *Dissert. Abstr.*, 1953, **13**, 902.
- MEAD, MARGARET. *Male and female*. New York: Morrow, 1949.
- MILLER, D. C. Morale of college-trained adults. *Amer. sociol. Rev.*, 1940, **5**, 880-889.
- MILLER, D. C. Economic factors in the morale of college-trained adults. *Amer. J. Sociol.*, 1941, **47**, 139-156.
- PECK, L. A study of the adjustment difficulties of a group of women teachers. *J. educ. Psychol.*, 1936, **27**, 401-416.
- PELZ, D. C. Influence: A key to effective leadership in the first-line supervisor. *Personnel*, 1952, **29**, 209-217.
- PETERSON, D. G., & STONE, C. H. Dissatisfaction with life work among adult workers. *Occupations*, 1942, **21**, 219-221.
- SMITH, MAY. The temperamental factor in industry. *Hum. Factors, London*, 1936, **10**, 301-314.
- SMITH, PATRICIA C. The prediction of individual differences in susceptibility to industrial monotony. *J. appl. Psychol.*, 1955, **39**, 322-329.
- SMITH, PATRICIA C. Cornell studies in methods of measuring job satisfaction: Introduction and scope. Paper read at American Psychological Association, New York, 1961.
- SMITH, PATRICIA C. Cornell studies of job satisfaction: I. Strategy for the development of a general theory of job satisfaction. Ithaca: Cornell University, 1963. (Mimeo)
- SMITH, PATRICIA C., & KENDALL, L. M. Cornell studies of job satisfaction: VI. Implications for the future. Ithaca: Cornell University, 1963. (Mimeo)
- STOCKFORD, L. O., & KUNZE, K. R. Psychology and the pay check. *Personnel*, 1950, **27**, 129-143.
- SURVEY RESEARCH CENTER. Effective morale. Part 3. *Fortune*, 1950, **42**(8), 46-50.
- TWERY, R., SCHMID, J., JR., & WRIGLEY, C. Some factors in job satisfaction: A comparison of three methods of analysis. *Educ. psychol. Measmt*, 1958, **18**, 189-202.
- VROOM, V., & MANN, F. C. Leader authoritarianism and employee attitudes. *Personnel Psychol.*, 1960, **13**, 125-140.
- WEITZ, J. A neglected concept in the study of job satisfaction. *Personnel Psychol.*, 1952, **5**, 201-205.
- WHERRY, R. J. An orthogonal re-rotation of the Baehr and Ash studies of the SRA employee inventory. *Personnel Psychol.*, 1954, **7**, 365-380.
- WHERRY, R. J. Factor analysis of morale data: Reliability and validity. *Personnel Psychol.*, 1958, **11**, 78-89.

(Received March 11, 1963)

BRAND LOYALTY REVISITED:

A TWENTY-YEAR REPORT¹

LESTER GUEST

Pennsylvania State University

Consistency of brand preferences and the relationship between preferences and use were investigated by mail questionnaires 20 years after original preferences were specified. Reasons for discrepancies between current preferences and use were also solicited. The relationships between consistencies of the 162 respondents and some standard personal characteristics were determined. Average amount of agreement for stated preferences separated by a 20-year span was 26%, and for preference in 1941 and use in 1961 was 23%. A small degree of relationship was found between intelligence and preference agreement and original socioeconomic status and preference agreement. There was no evidence of a general factor of loyalty.

During the spring of 1953, a follow-up study of brand preference and use for 15 products was conducted (Guest, 1955). Twelve years before, 813 public school students distributed across Grades 3-11 (last year of high school) had given evidences of brand awarenesses and preferences (Guest, 1942, 1944). In spite of obvious inadequacies, it was necessary to collect subsequent data by mail questionnaire. Usable responses in 1953 were returned from 165 respondents, and in 1961 from 162 respondents.

Although the term loyalty seems in disrepute in some quarters because of known brand switching, possibly resulting from the many variables operating on purchasers, it is nevertheless important to determine the degree of consistency of preferences and the degree of relationship between preferences and purchasing behavior. Cunningham (1956) presents some evidence on these questions obtained using other techniques and time intervals.

The present report relates results of a 20-year follow-up of the 1941 universe to determine the degree of correspondence of previous preferences with present preferences and with brand usage.

METHOD

As in 1953, questionnaires eliciting brand preference, brand use, and reasons for discrepancies when

they existed, were sent to all persons for whom any reasonable addresses could be determined. The questions concerning preference were in multiple-choice form with opportunity to indicate preference for one of five brands listed for each product or to indicate that none of those listed was preferred. If "None" was chosen, respondents were asked to indicate also which of those listed would be preferred. Each respondent was also asked which brand was owned or used most frequently. Finally, reasons for preference-use discrepancies were ascertained by means of a check list. Although there was reason to consider some changes in format in 1961, for consistency it was thought wiser to retain the same format that was previously used as nearly as possible.

Of the 424 questionnaires mailed (52% of the 1941 universe), 149 (35%) were returned for want of a better address (none was available), 104 (25%) apparently were received by a responsible person but were never returned, 9 (2%) were found to be addressed to deceased persons, and 162 (38%) were returned completed. This compares favorably with the 1953 results although it is far from satisfactory. It is interesting to note that of the 162 usable returns, 75 (46%) had *not* returned a questionnaire in 1953. Hence, comparisons between years should take into account that precisely the same subjects were not involved both times and that 1953-61 comparisons are based on a very small number of cases.

Table 1 presents the sample data from the present study with comparison data from the original sample and the 1953 sample of that sample. One should note the tendency (apparent in 1953) for distortion of mail returns compared with the original sample. Except for the variable of sex, the 1961 sample continues the trend noted in 1953 for the higher socioeconomic groups and the higher IQ groups to respond at a statistically significant (.01) higher rate than others. Except for family status, the classifications are those imposed in 1941 and so represent a status of 20 years previous. Some of these clearly would not change, some might show minimal change,

¹Funds for this research were made available by the Central Fund for Research of the Pennsylvania State University.

TABLE 1
SAMPLE DATA

Variable	1941	1953	1961
Sex			
Males	52	46	55
Females	48	54	45
<i>N</i>	813	165	162
Age			
7	0	1	1
8	8	9	15
9	10	10	12
10	10	9	12
11	9	11	7
12	9	7	7
13	10	7	7
14	11	9	9
15	13	19	14
16	11	12	10
17	5	4	6
18	3	2	1
<i>N</i>	813	165	162
Socioeconomic status			
A	5	8	12
B	38	45	49
C	44	39	30
D	9	4	5
?	4	4	4
<i>N</i>	813	165	162
IQ			
79 down	1	1	1
80-89	5	1	2
90-99	15	10	7
100-109	22	24	26
110-119	20	25	22
120-129	11	16	17
130 up	4	8	7
?	21	15	18
<i>N</i>	813	165	162
Family status			
Married (no children)	—	12	6
Married (1 child)	—	9	13
Married (2 children)	—	7	38
Married (3 children)	—	3	20
Married (4 children)	—	1	13
Unmarried	100	19	10
?	—	49	1
<i>N</i>	813	165	162

Note.—All data are given in percentages with the exception of *N*s.

and some, for example, socioeconomic status, might be radically different in 1961. Clearly caution should be exercised in generalizing from this small sample which also might be biased for the major factors under study. The data that follow should be

of some value when considered with appropriate reservations.

RESULTS

The responses permitted the assembling of the data in Table 2. The left half of the table is concerned with comparisons of brand preferences for various years. Columns 5, 6, 11, and 12 are similar to the others except that whenever a respondent indicated that the 1941 preference was for a brand not listed, his responses for that product category were excluded. The latter procedure was followed because lumping many unlisted brands together might lead to artificially high correspondences. As the results show, this did not seem to be important in most instances.

Although there are individual differences in amount of preference agreement among the 15 brand categories, ranges are generally not great. Considering average percentages, 1961 results provide a third point for a trend line. For example, degree of agreement with 1941 data gets smaller the farther one gets from the preferences made as a child. There was 32% agreement between 1941 and 1953 preferences, a drop to 26% between 1941 and 1961, but 44% agreement between 1953 and 1961 (note Columns 4 and 10). The shorter time interval plus the fact that most subjects had reached adult status and possibly answered somewhat more realistically in 1953 than in 1941 might explain the results. Almost the same results were obtained when the "none of those" category was omitted. After 20 years, about one quarter of the subjects who responded prefer the same brands of products preferred as school age young people.

Even more important is the relationship between preference and use, and as the data in the right half of the table show, nearly the same results as above were obtained. On the average, a little less than one fourth of the subjects say they use the brands in 1961 that they said they preferred in 1941, which is only slightly less than the percentage obtained for 1941-53 time period. These data seem unaffected by eliminating the "none-of-these" category. The data in Column 9 show that there is a very high correspondence between stated preferences and use when both sets of

TABLE 2
PERCENTAGES OF AGREEMENT BETWEEN VARIOUS YEARS

Category	Preferences with preferences						Preferences with use					
	Total sample				Minus "none listed"		Total sample				Minus "none listed"	
	1941- 53	1941- 61	1953- 61	Diff. (3-2)	1941- 53	1941- 61	1941- 53	1941- 61	1961- 61	Diff. (9-8)	1941- 53	1941- 61
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Coffee	35	31	37	6	33	26	33	35	73	38	27	29
Typewriter	36	26	54	28	38	28	25	23	41	18	25	21
Dept. store	38	33	62	29	38	33	36	26	49	23	36	27
Automobile	24	19	47	23	23	18	20	19	58	39	18	18
Gasoline	33	22	45	23	33	24	29	21	76	55	30	22
Razor	38	39	61	22	51	49	36	38	90	52	50	46
Magazine	36	24	41	17	35	24	32	21	70	49	32	21
Watch	35	29	50	21	35	30	26	27	61	34	22	21
Tooth paste	27	18	44	26	25	16	29	16	85	69	28	14
Soap	31	20	41	21	31	20	30	20	85	65	31	21
Cereal	23	26	46	20	21	25	25	23	80	57	22	22
Bread	31	20	30	10	30	18	23	14	73	59	22	13
Tire	34	29	24	-5	36	32	25	20	59	39	24	18
Gum	29	24	52	28	28	24	28	24	83	59	27	24
Radio	31	24	32	18	38	25	22	23	54	31	22	22
Average	32	26	44	18	32	26	28	23	69	46	27	23

data refer to 1961, although there is variation from 41% to 90% for individual products. It is thought that these results indicate a substantial amount of agreement and loyalty toward brands of products, especially when, as mentioned below, there are "legitimate" reasons for many lacks of correspondence. Consideration of what degree of agreement is high is partly judgmental. For instance, it might be argued that heavy concentrations of preferences or usages artificially increase degrees of agreement. It is true that these are not spread equally across brands or across products. However, there is nothing to prohibit changes in preference (and often, use). Furthermore, in the case of use, if one brand has a high share of the market so that shifts in and out of the market are *minimal*, this is only supporting evidence of strong loyalty obtained from data external to the present study. At any rate, inspection of the data indicates that in only four cases was either preference or use for any yearly comparisons concentrated in one category more than 50% of the time, and there were relatively few cases where the concentration exceeded 40%.

Whenever there was a discrepancy between 1961 preference and use, the subjects were asked to indicate the reason on a prepared check list. Such a procedure can result in rationalized and perfunctory answers, and

TABLE 3
REASONS FOR DISCREPANCIES BETWEEN 1961 PREFERENCE AND 1961 USE

Someone else decided on brand used; got as a gift	23.60%
Difference in price overrides preference	16.13
Don't use, don't own, don't buy	14.87
Not sold around here, no such store, just can't get	7.47
Inconvenience in getting brand or to place where available	7.00
Have no real preference, just don't care	3.47
Bought to please someone else, husband, wife, etc.	3.33
Haven't yet had chance to change to preferred brand	2.53
Difference in quality	2.33
Others (less than 2% each): dealer relationships (likes or dislikes, personal friend or relative), get more prestige where friends buy or go, special premiums, just likes variety, has charge account, don't know	4.87
Confused answers and reasons	6.73
No answer-omission	7.07
Total	99.40

TABLE 4

AVERAGE PERCENTAGE AGREEMENT BETWEEN 1941-61
PREFERENCES BY STANDARD VARIABLES

Variable	Obtained	Expected
Sex		
Males	55	55
Females	45	45
Socioeconomic status		
High	68	64
Low	32	36
Age		
Young	33	40
Middle	32	30
Old	34	31
Intelligence		
High	30	30
Middle	58	58
Low	12	12
Marital status (1952)		
Married (few children)	54	55
Married (many children)	10	7
Unmarried	36	37
Marital status (1961)		
Married (few children)	57	57
Married (many children)	32	33
Unmarried	10	11

even carefully probed personal interviewing realizes difficulties isolating “real” motivations. However, the data presented in Table 3 do permit some tentative conclusions.

The data presented are averaged percentages using small and variable *N*s (those where there is disagreement between 1961 preference and use; *N*s between 16 and 96). The primary reasons for discrepancies vary a great deal among some product categories. The fourth and fifth reasons listed seem especially legitimate because some of the brands listed in 1941 no longer exist (Packard, Collier’s) or are local brand names that subjects, distant from their original homes, cannot buy. Most discrepancies are for “good” reasons, at least half of them being “unavoidable.” Therefore, in the absence of special limiting conditions, stated preferences provide pretty fair indications of likely purchasing behavior.

A number of variables often related to attitudes that are routinely checked were analyzed in relation to preferences given in 1941 and 1961. Again, average percentages were computed with no attempt to alter the automatic weights from obtained numbers of cases.

The data in Table 4 vary to some degree with those found in 1953. Here, the differences between the obtained percentages and the expected percentages in terms of size of the group are negligible with one or two exceptions. It does appear that more of the high economic group than the low have loyalty in terms of expectation, and that slightly more of the older subjects have loyalty. In both cases the differences are small, even though the former is significant at the 5% level and the latter significant at the 1% level. Some of the comparisons by individual products result in differences, but they are not large as a rule, nor do they seem to be consistent with expectations (e.g., higher socioeconomic group more likely to have higher agreements for expensive products such as automobiles or watches). In the light of these data, it would not seem as if degree of loyalty is a function of these variables.

Finally, it was thought important to ascertain the degree of loyalty within subjects across brands to determine whether there might be some persons who were generally loyal, and others for whom there was little or no loyalty. For each subject, there were 15 opportunities to display loyalty, both for 1941-61 preferences, and for 1941-61 preferences and use. The results indicate that loyalty is not a general characteristic of persons but rather of brands and products, and these are different for different persons. Not many subjects have few or no agreements for the 15 categories, and few have loyalties for brands for over half the products. For the 1941-61 preference comparison, the median number of agreements is 3.53 (mode = 3, range from 0-9), and for the 1941 preference and 1961 use comparison, the median number of agreements is 3.13 (mode = 3, range from 0-8). In spite of the inability to be consistent in some cases because of the demise or unavailability of some brands, the number of

agreements per person is relatively high, especially with use. It should be remembered that these subjects could hardly have inflated their agreements purposely since they would not likely have remembered what they said in 1941 in response to the questions.

CONCLUSIONS

It is surprising how much correspondence exists between the results of the 1953 study and the present follow up. Hence, these conclusions are almost replications of those reached earlier.

This 20-year follow up presents suggestive evidence that there is a rather high degree of loyalty toward brand names, especially where special considerations such as unavailability, price considerations, and the respondent not being the primary purchaser, do not play a major part in brand selection. After a lapse of 20 years, with preferences originally being verbally expressed during the ages of 7 through 18, there is an average amount of agreement between early and current preferences of 26%. The average degree of agreement between 1941 preferences and 1961 use is slightly less, 23%.

In the present study there is little indication that sex, intelligence, or marital status is

related to preference agreements, and only slight indication that a few more subjects in the higher socioeconomic status groups have greater preference agreements. Opposed to the 1953 results, it does appear that the older subjects have more agreements than the younger ones, perhaps because original preferences might have been more rational. There still is no indication that there is a general loyalty factor in people. Rather it appears that it is related to specific products and brands, as well as to special pressures impinging upon people. Although most people can be changed and do change over a long period of time, the results of this study show that even early childhood experiences exert considerable influence upon later brand purchasing behavior. The implications for other preference areas should be clear.

REFERENCES

- CUNNINGHAM, R. Brand loyalty: What, where, how much? *Harv. Bus. Rev.*, 1956, **34**, 116-128.
GUEST, L. The genesis of brand awareness. *J. appl. Psychol.*, 1942, **26**, 800-808.
GUEST, L. A study of brand loyalty. *J. appl. Psychol.*, 1944, **28**, 16-27.
GUEST, L. Brand loyalty: Twelve years later. *J. appl. Psychol.*, 1955, **39**, 405-408.

(Received March 25, 1963)

TEST DIFFICULTY, RELIABILITY, AND DISCRIMINATION AS FUNCTIONS OF ITEM DIFFICULTY ORDER

MARSHALL H. BRENNER¹

Ohio State University

In an attempt to evaluate the often stated rule that test items should be arranged in an increasing order of difficulty, the effect of item difficulty order on total test reliability, difficulty, and discrimination was investigated in a series of 4 experiments. Each experiment involved a comparison of 2 or more tests, containing the same 40 items and differing only with respect to the order of those items. The differently ordered forms did not, in any of the experiments, differ significantly in test difficulty or test reliability. The results with respect to discrimination were not as clear-cut. However, the results tend to lead to the conclusion that item difficulty order on a power test of facts and principles given in the normal college classroom will not significantly affect these 3 test statistics.

For years it has been stated, with little or no empirical evidence, that test items should be arranged in order of increasing difficulty; that is, with the easiest item first, the next easiest item second, etc., with the hardest item placed last (Remmers & Gage, 1943; Ross, 1947; Ruch, 1929). The little experimentation that has been done that is related to this subject indicates that such an increasing order results in higher total test scores (Lund, 1953), particularly in the case of speeded tests. But no attempt has yet been made to determine the effect of item order on the discriminative power of the test, even though discrimination is one of the foremost values of any test being used for evaluation. The study described here was undertaken to evaluate the effect of different item difficulty orders on test discrimination, reliability, and difficulty. The null hypothesis was that no order would be significantly different than any other order for these three statistics.

METHOD

The study covered a period of two quarters, the winter and spring quarters of the 1961-62 academic year, and was done in the course Educational Psychology 407 at the Ohio State University. This course is required of all education majors and covers the major areas of educational psychology. Four "midterm" tests were administered throughout each quarter, each test covering one of the major topical areas in the course: development, principles

of learning, applied learning, and measurement and remediation.

For each of the four midterm tests given in the winter quarter, two 40-item tests were constructed. All the items were original ones constructed by the investigator, and all were of the multiple-choice format with four alternatives. The items, and the correct alternatives, were arranged randomly through the use of a random numbers table.

The tests were administered to all sections by their regular instructors at the normal class hour on the third, fifth, seventh, and ninth Mondays of the quarter. The students were given 50 minutes to complete the tests. Fewer than 1% of the students indicated that they had not had a chance to consider each item thoroughly; the tests are, therefore, considered to have been power tests with little or no time factor involved.

The difficulty (percentage passing, uncorrected for chance) and the discrimination value (point-biserial correlation between success on the item and total test score) were computed for each of the 320 items used in the quarter.

In the spring quarter, the items were arranged in a number of experimental orders on the basis of the item difficulties obtained in the winter quarter. The same chapters were tested, under the same conditions, on the same relative days, as in the winter quarter. The four test administrations are referred to below as Experiments I, II, III, and IV. In each of the four experiments, the same 40 items were used on all of the forms used in that experiment; the only difference between forms was in the ordering of the items.

Experiment I

The procedure of evaluating the effect of the ordering of items on one test by using the difficulty values obtained on a previous administration rests on the assumption that the difficulty values for individual items consistently retain their relative order;

¹ Now at the Lockheed-California Company, Burbank.

that is, that item difficulties are reliable. Previous investigators have found item difficulties to be highly reliable (Carter, 1942; Davis, 1951; Gibbons, 1940). Rather than rely completely on other investigations, however, an empirical check was made on the reliability in this situation by readministering one of the two forms given on the first midterm of the winter quarter exactly as the items were ordered there (in a random order of difficulty). This test was denoted as Form IB.

These same 40 items were also arranged in two other orders for administration at the same time: (a) an increasing order of difficulty form, in which the items were arranged in a progressive order of difficulty, the easiest item first and the hardest item last (denoted Form IA) and (b) a decreasing order of difficulty form, in which the items were arranged in the exact reverse of the increasing order, with the hardest item first and the easiest item last (denoted as Form IC).

Experiment II

Of the 80 items written for Midterm 2 of the winter quarter, 40 were selected for use in Midterm 2 of the spring quarter. The selection was done in the following manner: First, the 10 easiest and the 10 most difficult items were selected. The other 20 items were then selected in such a manner that each of the three chapters tested was given approximately equal weight, the most important principles in the area being tested were included, items from all levels of difficulty were selected, and the better discrimination values were selected whenever a choice was possible.

Two forms were constructed with these 40 items: Form IIA, in which the 10 easiest items were placed at the beginning, in increasing difficulty, and the other 30 items were randomly arranged, and Form IIB, in which the 10 hardest items, in decreasing difficulty order, were placed at the beginning and the other 30 items were randomly arranged.

Experiment III

The 40 items used on one of the two midterm tests given as the third midterm of the winter quarter were rearranged into two differently ordered forms and administered as the third midterm test of the spring quarter. The items of Form IIIA were arranged in an increasing order of difficulty and the items of Form IIIB were arranged in a decreasing difficulty order.

Experiment IV

The same procedure used to select the 40 items used in Experiment II was also employed in Experiment IV. The items were then arranged into two forms: Form IVA, an increasing difficulty order, and Form IVB, a decreasing difficulty order.

RESULTS

Three measures were obtained for each of the experimental forms: Difficulty (average

TABLE 1

TEST DIFFICULTIES, RELIABILITIES, DISCRIMINATION VALUES, AND SIGNIFICANCE TESTS

Form	Difficulty	Reliability	Discrimination
IA	21.18	.578	.220
IB	21.04 $p > .50$.553 $p > .75$.232 $p > .40$
IC	20.90	.598	.218
IIA	24.04	.753	.283
IIB	23.93 $p > .70$.674 $p > .40$.250 $p > .02$
IIIA	26.14	.778	.309
IIIB	26.33 $p > .60$.805 $p > .70$.326 $p > .20$
IVA	23.69	.736	.284
IVB	24.17 $p > .30$.747 $p > .90$.289 $p > .70$

number of items answered correctly), Reliability (as evaluated by the Kuder-Richardson Formula 8), and Discrimination (average point-biserial correlation between item and total test score). These results are presented in Table 1.

For each experiment, significance tests were computed for the difference between the forms for each of these measures. In all cases, significance was evaluated by use of the appropriate t test. The probabilities obtained from these tests are also presented in Table 1. In that table, however, only one probability is reported for each measure in Experiment I; the probability stated corresponds to the largest difference obtained between any two of the forms.

Difficulty

In none of the four experiments was there any difference large enough to have been obtained less than 30% of the time due to chance alone.

Reliability

In none of the four experiments was there a difference large enough to have been obtained less than 20% of the time due to chance alone.

Discrimination

In three of the experiments, none of the differences in discrimination was large enough

to have been obtained less than 40% of the time by chance alone. In the other experiment (Experiment II), the difference was significant between the .05 and .02 levels of significance.

DISCUSSION

The tests employed in this study varied from those normally used in the classroom situation in two respects: (a) the average score obtained was relatively low, ranging from 52% to 66% correct answers; (b) the range of item difficulties was consistently quite large, the smallest range being from .17 to .94. The effect of these two differences would be to increase the effect that item order might have on the test statistics obtained in this study. Insofar as no significant differences were obtained for test difficulty and test reliability, and only one of four differences in discrimination was significant at the .05 level (with two of the remaining three differences being in the opposite direction), it would appear that there is no value, for the average

college instructor, in spending the time and effort necessary to arrange the items in order of item difficulty when constructing a power test to measure the attainment of facts and principles from a text.

REFERENCES

- CARTER, H. How reliable are the common measures of difficulty and validity of objective test items? *J. Psychol.*, 1942, 13, 31-39.
- DAVIS, F. B. Item selection techniques. In E. F. Lindquist (Ed.), *Educational measurement*. Washington, D. C.: American Council on Education, 1951. Pp. 266-328.
- GIBBONS, C. C. The predictive value of the most valid items of an examination. *J. educ. Psychol.*, 1940, 31, 616-621.
- LUND, K. Test performance as related to order of item difficulty, anxiety, and intelligence. Unpublished doctoral dissertation, Northwestern University, 1953.
- REMMERS, H. H., & GAGE, N. L. *Educational measurement and evaluation*. New York: Harper, 1943.
- ROSS, C. C. *Measurement in today's schools*. (2nd ed.) New York: Prentice-Hall, 1947.
- RUCH, G. M. *The objective or new-type examination*. New York: Scott, Foresman, 1929.

(Received March 25, 1963)

LIFE HISTORY ANTECEDENTS OF SALES, RESEARCH, AND GENERAL ENGINEERING INTEREST

FREDERICK B. CHANEY¹ AND WILLIAM A. OWENS

Purdue University

Responses to a 170-item, multiple-choice life history questionnaire were analyzed against criteria of sales, research, and general engineering interest. A sample of 388 university freshmen in engineering was used for the sales and research portions of the analysis and 700 freshmen were used in appraising general engineering interest. The significant items were summarized by content category and used to develop scoring keys for each interest criterion. When these keys were cross-validated on completely independent samples, correlations of .57, .42, and .51 were obtained for sales, research, and general engineering, respectively. These findings provide empirical data regarding the development of engineering interest and suggest the potential feasibility of using scored life history data to predict the subsequent development of vocational interests.

Each year industry recruits thousands of graduates from our engineering schools and places them on a wide variety of specific job assignments. A better understanding of the interests of these new engineers could contribute to the accuracy of this personnel placement procedure.

Clark (1961) has recently demonstrated the importance of interest measures in predicting achievement and occupational choice for a number of nonprofessional vocations. In addition, the ability to predict occupational survival from inventoried interests has been clearly established by Strong (1957). With specific reference to tenure of engineering personnel, Boyd (1961) has successfully used the Strong Vocational Interest Blank (SVIB) to predict turnover at an electrical company.

While psychometricians have been reasonably successful in measuring general engineering interest, unfortunately very little has been done to increase our understanding of the developmental process. The work of Kulberg and Owens (1960) appears to be the only previous empirical investigation of the development and maturation of engineering interest. They correlated the life history responses of 111 mechanical engineering freshmen with scores obtained from the SVIB Engineer scale and special sales and research

engineering interest keys developed by Dunnette (1957). The present study is primarily an attempt to extend the validity of Kulberg's findings using a larger sample, more extensive questionnaire, and appropriate cross-validation procedures.

METHOD

Subjects. The subjects for this study consisted of 508 freshmen in the School of Engineering and 392 freshmen from the School of Agriculture at Purdue University. The engineering group included 90 "exceptional students" enrolled in an accelerated curriculum, 181 electrical engineers, 158 mechanical engineers, and 79 industrial engineers. The agriculture students could not be classified by major areas since their school does not allow specialization until the sophomore year. This total sample of 900 individuals represents 93% of the students originally tested. The entire sample was used for the general engineering interest analysis, while only the 508 engineering students were used in the sales and research analyses. It is important to realize that inventoried interests and not curricular groups were used as criteria in the analysis. The agriculture students were included primarily to eliminate restriction of range at the lower end of the General Engineering Interest scale. These individuals, rather than some other curricular group, were used because both criterion and life history data on the agriculture students were available from another study at the time the analysis was performed. The sample selection was further justified by the nature of the resulting criterion distributions.

Criteria. The criterion of general engineering interest was the Engineer scale on the SVIB. The mean scale scores were 36.52 and 31.54 for the engineering and agriculture subsamples. The distribution for the total sample was approximately normal

¹ Currently a NATO Postdoctoral Fellow in Science, University of London.

with a mean of 34.35 and a standard deviation of 9.48. The Engineer scale scores were dichotomized at the median (35.08) for the item analysis. The criteria of sales and research engineering interest were obtained from special profile keys developed by Dunnnette (1957). These distributions were also dichotomized at their medians which were 25.50 and 34.03 for Sales and Research scales, respectively. The test-retest reliability of the Engineer scale over a 1-year period for 247 college freshmen is .91 (Darley, 1955) and a coefficient of internal consistency of .94 has been reported by Strong (1959). Since no reliability data were available for the Dunnnette keys, internal consistency estimates were computed by dividing the keys on an odd-even basis. The resulting estimates, corrected for length by applying the Spearman-Brown Prophecy Formula, were .68 for the research key and .87 for the sales key.

Predictors. The predictor items for the analysis consisted of 170 life history items drawn from the following sources: 57 from Kulberg and Owens (1960); 39 from Smith, Albright, Glennon, and Owens (1961); and 74 from the co-author's personal file over 1,000 life history items. These items were intended to sample the entire range of experiences which could conceivably be related to the development of engineering interest. The specific content dealt with childhood and early experiences, academic achievement, occupational status of parents, sibling relationships, socioeconomic conditions, habits, attitudes, friends, and other background information.

Item Analysis. The item analysis employing general engineering interest as a criterion utilized 700 of the 900 subjects with 200 individuals being held out to estimate the concurrent validity of the selected items. The 700 subjects were randomly divided into Subsamples A and B with an equal distribution of high and low scores. Using the program developed by Smith et al. (1961) for the IBM 705, the significance of the difference between response percentages of high and low criterion groups was computed for each response option for Subsamples A and B. Those options which were significant beyond the .20 level in both A and B were then combined via Baker's (1952) Table for compound probabilities to provide a single estimate of significance based upon an N of 700.

The procedure outlined above was repeated with the sales and research engineering interest criteria. This item analysis used 388 of the 508 engineering students with the remaining 120 engineers being retained for cross-validation of the selected items. Once again, compound probabilities were obtained by way of Baker's Table for each item that was significant beyond the .20 level in both subsamples.

Cross-Validation. Before constructing life history keys for each of the three criteria, the options selected by less than 5% of the sample were eliminated from consideration. The remaining items with a compound probability of $p < .05$ were given weights of +1 or -1 in accordance with the direction of discrimination. The possibility of using

differential weighting was rejected on the basis of Guilford's (1954) conclusion that differential weighting pays small dividends when there are over 10-20 items to be combined.

The life history key for general engineering interest was applied to the 200 individuals who were excluded from the item analysis sample and the resulting composite scores were correlated with the SVIB Engineer scores. This procedure provided an unbiased estimate of the concurrent validity of the life history composite. The same cross-validation procedure was followed with sales and research engineering life history keys, using the remaining 120 subjects in the engineering sample.

RESULTS

The significant options for each criterion were placed in one of the following content categories: Academic, Family and Community, Interpersonal Relations, Occupational, Recreation and Activities, and Self-Perception. The content of the items which yielded a compound probability of $p < .05$ is summarized below for each of the three interest criteria.

General Engineering Interest. The significant academic life history items indicated that individuals with a high degree of general engineering interest came from relatively larger schools. They elected, enjoyed, and did their best work in physical science and mathematics courses. They disliked and had their most difficulty with courses in history, economics, and civics. Natural science, biological science and commercial courses were also difficult for these individuals. These people received a great deal of enjoyment from problem solving courses and tended to be quite persistent in this type of activity. They disliked discussion courses and avoided situations requiring a high level of verbal skill. The individuals in the high criterion group for general engineering interest tend to come from upper-middle class sections of towns whose populations are between 5,000 and 25,000. Their parents tend to be rather inactive socially and would like the student to enter a profession.

The significant interpersonal relations items indicate that general engineering interest is negatively related to frequency of dating and 14% of the subjects had not started dating by their freshman year in college. Individuals with a high degree of gen-

eral engineering interest do not feel free to express their views to their associates, seldom tell their troubles to other people, tend to have close friends, and are somewhat uneasy about meeting new people. These individuals seek positions that will allow them to develop and use their own ideas. They prefer research or design activities in small group situations which provide a minimum of interaction with others.

Responses covering recreation and activities indicate that persons with a high degree of general engineering interest tend to avoid high school activities or politics and that they derive most of their satisfaction from academic success. They do most of their leisure reading in technical books and prefer science fiction to other radio, TV, and magazine subject matter. These individuals describe themselves as being filled with curiosity and somewhat more creative than average. They are very confident about their intellectual ability but not as confident about their social skills.

Sales Engineering Interest. The few academic items which were predictive of the sales criterion indicated that this type of interest is positively related to a liking for school in general and negatively related to enjoyment of natural and biological science. These individuals were only average students in science and math courses.

Persons with a high degree of sales interest tend to come from the exclusive sections of cities over 25,000. They characterize their fathers as strict, moralistic individuals employed in business, sales, or supervisory work. These individuals were given the freedom to spend their evenings away from home during their high school years and reportedly had more money to spend than their classmates. Their parents constantly impressed upon them the need for a good education.

In the interpersonal area, sales interest appears to be associated with early and frequent dating. Individuals with a high degree of sales interest frequently express their views to others and report that they have considerable impact on their social group. They meet people easily and claim to have many close friends.

Persons high in sales interest desire a

position with authority and responsibility in an occupation which allows a great deal of interaction with other people. These individuals participated frequently in high school activities and were quite successful in school politics. They have had a wide variety of experiences and have organized a number of different activities. They also perceive themselves as being more popular than others and feel that religion is an important factor in their lives.

Research Interest. Persons measuring high in research interest enjoyed math and physical science courses and tended to be persistent problem solvers. They reportedly were not allowed out in the evenings during high school and their fathers were not in business or sales work.

The high criterion group for research interest contained significantly more individuals who had not started dating than the low group. Persons high in research interest reported that in choosing for games, they were chosen near the end and that they seldom tell others their troubles. One builds an image of a very reserved individual who finds it somewhat difficult to relate to others. As would be expected, these persons show a high preference for research positions which will enable them to work with ideas and experiment with new methods.

Research interest was consistently found to be negatively related to participating in high school and recreational activities. On the other hand, these individuals did report a liking of operas, symphonies, and concerts. They do not view themselves as being particularly popular or self-confident and they feel that professional status and authority are very important.

The number of significant items (compound $p < .05$ on one or more options) for each criterion is summarized by content category in Table 1. The academic and occupational categories produced the largest proportion of significant items for the general engineering interest criterion with the Self-Perception and Family and Community categories being the least productive. The Occupational category also gave the highest proportion of significant items for the sales and research engineering interest criterion.

TABLE 1

CLASSIFICATION OF SIGNIFICANT ITEMS

Content category	Total number of items	Criterion					
		Engineering		Sales		Research	
		<i>n</i>	<i>p</i>	<i>n</i>	<i>p</i>	<i>n</i>	<i>p</i>
Academic	23	15	.65	4	.17	2	.09
Family & Community	32	7	.22	8	.25	2	.06
Interpersonal Relations	20	5	.25	5	.25	2	.10
Occupational	18	8	.44	7	.39	5	.28
Recreation & Activities	34	11	.32	10	.29	6	.18
Self-Perception	43	10	.23	9	.21	5	.12
Total	170	56		43		22	

Note.—*n* = number of significant items; *p* = proportion of significant items.

Validity and Reliability of Scoring Keys. Product-moment correlations were computed between the composite life history scores and the three interest criteria. None of the subjects used in this cross-validation were used in the item analysis. The resulting concurrent validity coefficients were .51, .57, and .42 for general, sales, and research engineering interest, respectively. All of the correlations were significant beyond the .01 level.

To obtain a reliability estimate ² for some of the scored life history data, the 56-item form was readministered after a period of 19 months had elapsed. Based upon the results for 49 subjects, a test-retest coefficient of .85 was obtained. This value is somewhat higher than the .68 reported by Smith et al. (1961) for a 29-item key which was re-administered after 2 months.

DISCUSSION

A comparison of the life history antecedents of sales, research, and engineering interest which were identified in the present investigation strongly supports Kulberg's conclusions concerning the differential antecedents of these several types of interest. He summarized his findings by stating that engineering interest is preceded by a history of academic superiority and by more satisfactory experiences with things and ideas than with people and social situations. This

pattern appeared to be intensified in the case of research interest and reversed for sales engineering interest. While a great deal of time could be spent discussing similarities and differences for individual items, it would be hard to improve on this overall synthesis of the data.

It is interesting to note, however, that none of the items which are common to both general and research engineering interest discriminate in opposite directions; there were simply many more antecedents of engineering interest. This fact might lead one to view research interest as a characteristic that develops from the same types of experiences as general engineering interest, with its unfolding being delayed until late adolescence or early adulthood. On the other hand, differential antecedents for sales and general engineering interest appear at quite an early age. Even the earliest antecedents of these two patterns are negatively related and this relationship is maintained throughout the developmental period. Thus sales and general engineering interest appear to be at opposite poles of an interest continuum from their very beginning.

Perhaps the most obvious criticism of this research would be that it simply demonstrates a correlation between two sets of verbal self-reports. In defense, one can only say that the use of inventoried interests as an intermediate criterion permitted a necessary first step toward a better understanding of interest development. In addition,

² The writers wish to acknowledge the assistance of Alan R. Starry in conducting the reliability study.

the wealth of data on the SVIB scales suggest that they are highly relevant to "more ultimate" criteria of vocational interests. It is also important to realize that life history data are generally backed by firsthand experience at a much earlier age than interest inventory responses.

It is difficult to generalize from these findings since all the subjects were freshmen from a single university. However, the synthesized life history antecedents do present an empirically derived picture of selected aspects of engineering interest development. Taken as a scored composite, they also offer quite convincing evidence for the potential value of life history data in predicting the general nature of subsequent vocational interests. These results suggest the possibility of developing a measure of vocational interest, responses to which would stabilize much earlier than those of present inventories. Such an instrument would be based on actual life history data rather than on preferences for activities which might be out of the subject's range of experience.

While the use of traditional interest inventories has met resistance in most technical employment situations, life history data may be relatively easily collected via application blanks or standardized interviews. The foregoing findings suggest that the scored life his-

tory technique may be of value in the placement of engineering personnel and possibly also in their selection.

REFERENCES

- BAKER, P. C. Combining tests of significance in cross-validation. *Educ. psychol. Measmt.*, 1952, **12**, 300-306.
- BOYD, J. B. Interests of engineers related to turnover, selection, and management. *J. appl. Psychol.*, 1961, **45**, 143-149.
- CLARK, K. E. *Vocational interests of nonprofessional men*. Minneapolis: Univer. Minnesota Press, 1961.
- DARLEY, J. G., & HAGENAH, THEDA. *Vocational interest measurement: Theory and practice*. Minneapolis: Univer. Minnesota Press, 1955.
- DUNNETTE, M. D. Vocational interest differences among engineers employed in different functions. *J. appl. Psychol.*, 1957, **41**, 273-278.
- GUILFORD, J. P. *Psychometric methods*. New York: McGraw-Hill, 1954.
- KULBERG, G. E., & OWENS, W. A. Some life history antecedents of engineering interests. *J. educ. Psychol.*, 1960, **51**, 26-31.
- SMITH, W. J., ALBRIGHT, L. E., GLENNON, J. R., & OWENS, W. A. The prediction of research competence and creativity from personal history. *J. appl. Psychol.*, 1961, **45**, 59-62.
- STRONG, E. K., JR. Prediction of educational and vocational success through interest measurement. In, *Proceedings of the 1957 Invitational Conference on Testing Problems*. Princeton, N. J.: Educational Testing Service, 1957. Pp. 72-82.
- STRONG, E. K., JR. *Manual for Strong Vocational Interest Blank for men and women*. Palo Alto, Calif.: Consulting Psychologists Press, 1959.

(Received March 4, 1963)

SUPERVISOR ESTEEM AND PERSONNEL EVALUATIONS¹

PAUL D. NELSON

United States Navy Medical Neuropsychiatric Research Unit, San Diego, California

Personnel evaluations given by 7 pairs of Antarctic station supervisors were analyzed to determine the extent to which such judgments can be affected by the level of esteem at which a supervisor is held. Agreement between supervisors in their evaluations of personnel was not related to the degree of mutual esteem between supervisors and, similarly, the esteem held for supervisors by their subordinate peer group was not related to supervisor-peer agreement in evaluating personnel. Agreement between supervisors was related ($p < .05$) to the level of esteem held for them by the peer group. Finally, supervisors of different levels of esteem seemed to place similar value on various performance characteristics.

In any organization within which supervisor evaluations of personnel serve as criteria of individual performance it is crucial to know something of the validity of such evaluations. If agreement between two or more supervisors contributes to the validity of personnel evaluations, it is important to understand the conditions or variables associated with supervisor agreement. One such variable might be the general level of esteem with which the supervisor is held by other supervisors and by the subordinate peer group being evaluated. Is the esteem with which a supervisor is held in any way related to the way in which he evaluates subordinate personnel?

The setting for the present study is the small Antarctic scientific station. Each year approximately 10 Navy enlisted personnel and 10 civilian scientists or technicians work and live together for 12 continuous months at each of three small stations. The groups are isolated from other groups, except for occasional radio communication, for at least 6 months of this time. To provide performance-criteria information for the screening and selection program, the two station leaders or supervisors, one Navy officer and one civilian, evaluate all station personnel on

specific and general performance dimensions. Although evaluations from the subordinate peer group are also obtained the limited success in obtaining the latter has rendered the supervisor evaluation an important single source of criterion information. The extent to which supervisors agree in their evaluations of personnel has therefore been of practical concern.

Three questions were posed in the present study. First, to what extent is the level of mutual esteem between supervisors related to their agreement in evaluating personnel? Secondly, to what extent is the esteem held for supervisors by the subordinate peer group related to both the agreement among supervisors and the agreement between supervisors and peers in their evaluations of personnel? Finally, do relatively high- and low-esteem supervisors value the same traits or behavior characteristics when they evaluate personnel?

METHOD

Subjects

The supervisors, who are the subjects in the present study, were seven military and seven civilian leaders from seven small Antarctic station groups, the latter ranging in size from 16 to 35 personnel. The median age for military supervisors was close to 28 years while that for civilian supervisors was approximately 35 years. All had at least a college education; most had more.

Procedures

At the end of the Antarctic year, the two supervisors from each station independently ranked all

¹The present study was supported by the Bureau of Medicine and Surgery, Department of the Navy, under Research Task MR005.12-2004, Subtask 1. The opinions expressed in this paper are those of the author and are not to be construed as necessarily reflecting the views or endorsement of the Navy Department.

personnel, including one another, in the order in which they would select them for wintering-over duty were they to return to the Antarctic. Station peers, either during or after the Antarctic year, were also asked to evaluate one another and the two supervisors on the same criterion item through either a ranking or nomination technique.² A final rank order of station personnel, including the supervisors, was then obtained from each supervisor separately and from the average of the station peers' evaluations. All ranks were converted to T scores ($M = 50$, $SD = 10$) for further analysis.

In addition to serving as the general criterion of personnel performance, the rankings on "want to return with" served as the source for estimates of supervisor esteem. Levels of mutual esteem between supervisors were estimated by the average of the T scores assigned the two supervisors at each station by one another. The indices of esteem held for supervisors by the subordinate peer group were obtained from both the average and range of T scores assigned the supervisors by the peer group in their evaluation of station personnel.³ Once the esteem estimates for supervisors were obtained, the supervisors were removed from all rankings and the remaining station personnel were reranked and assigned corresponding T scores again. In addition to the peer evaluations of station personnel and the individual supervisor evaluations, the two supervisors' evaluations at each station were combined to yield a joint supervisor evaluation. These data then served as the basis for analyses of agreement between supervisors themselves and between peers and supervisors.

Finally, the supervisors, in addition to the ranking task, rated all station personnel on the following behavior characteristics using a 9-point graphic scale: emotional control, acceptance of authority, self-confidence, achievement motivation, likeability by group members, alertness, industriousness, motivation to be an efficient group member, attitude towards the Antarctic project, and job assignment satisfaction. On the assumption that the magnitude of correlation between a specific behavior characteristic and the criterion item of "return with" is indicative of the relative value of such a characteristic to the supervisor, the relationships between the preceding ratings and the supervisor's ranking criterion were evaluated to determine the extent to which high- and low-esteem supervisors valued the same characteristics in ranking their men.

² The peer evaluation task varied for groups from different years.

³ Since there are slight variations in the upper and lower limits of T scores for different sized groups, all estimates of supervisor esteem were converted to ratio scores so as to indicate the average level or the range of esteem actually obtained with reference to the maximum average or range of esteem possible within each group.

RESULTS⁴

Measures of personnel-evaluation agreement were obtained by Pearson r applied to T scores from the criterion item of "return with." Agreement between supervisors themselves was of a relatively restricted range; the highest level of agreement ($r = .58$, $N = 14$) was not significantly different from the lowest level of agreement ($r = .33$, $N = 33$), using a test of difference between independently obtained correlation values (McNemar, 1955). Somewhat greater range in correlation values was observed for the agreement between the combined supervisor evaluation and the peer-group evaluation of personnel. Nevertheless, the values were statistically similar with only the lowest value ($r = .31$, $N = 33$) being significantly different from the highest value ($r = .84$, $N = 17$), $p < .01$. The median level of agreement between peers and the combined supervisors ($r = .65$) was somewhat higher than that between supervisors themselves ($r = .50$).

Although each individual supervisor's agreement with the peer-group evaluation was also obtained, only the combined supervisor evaluation was presently used in comparison with the peer evaluation for two reasons. First, for only one of the seven pairs of supervisors was there a significant ($p < .10$) difference between the two supervisors' levels of agreement with the peer-group evaluation using a test of difference between nonindependent correlation values (McNemar, 1955). Secondly, if the individual supervisor-peer agreement values were to be related to individual supervisor-esteem levels, the 14 possible pairs of such scores could not be treated as independent data.

The esteem- and personnel-evaluation agreement variables were then ranked for the seven stations. Rank-order correlations (ρ)

⁴ Data from which the results were obtained are contained in Tables A, B, and C deposited with the American Documentation Institute. Order Document No. 7824 from ADI Auxiliary Publications Project, Photoduplication Service, Library of Congress, Washington, D. C. 20540. Remit in advance \$1.25 for microfilm or \$1.25 for photocopies and make checks payable to: Chief, Photoduplication Service, Library of Congress.

were obtained among all variables. Only two of the relationships were significant within at least the .10 level of confidence, these being (a) the average level of esteem held for both supervisors by the peer group and its relation to the level of agreement between supervisors and (b) the correlation between station size and agreement between supervisors, the latter being in a negative direction. By employing a partial correlation to control for group size, the relationship between peer esteem for supervisors and agreement between supervisors was increased from $r' = .72$ to $r' = .85$ ($N = 7$, $p < .05$). No other relationship was increased to a significant level by partial correlation.

The results indicated moderate agreement ($r' = .63$) between the two estimates of supervisor esteem, one being derived from supervisors themselves and the other from peers. Mutual esteem between supervisors was not related to supervisor agreement in evaluating personnel; and, none of the indices of esteem were significantly related to the level of agreement between peers and supervisors.

The final analysis was that of assessing the similarities and differences between supervisors or relatively high and low esteem in terms of the value which they attached to specific behavior characteristics when making their rank-criterion judgments. In this analysis seven supervisors with esteem T scores ranging from 54 to 70 (median T score = 64) were compared with six supervisors with esteem T scores ranging from 34 to 44 (median T score = 38).⁵ For each individual supervisor, the ratings given to personnel on the specific behavior characteristics (i.e., emotional control, acceptance of authority, self-confidence) were correlated with the supervisor's ranking (T scores) of personnel on "want to return with." Due to the discontinuous nature of many of the sets of ratings, personnel were divided as close to the median as possible on each characteristic and biserial correlations were computed. The resulting correlations were then ranked from high to low for each supervisor to establish

the relative value attributed to each characteristic.

Among the relatively low-esteem supervisors there was significant agreement in the rank order of the characteristics in terms of their relationship with the criterion ($W_c = .521$, $p < .01$). No such agreement was observed among the supervisors of high esteem. For the high- and low-esteem groups separately an average-rank position was obtained for each characteristic and a final rerank established to compare the two esteem groups. The two final sets of ranks were quite similar as evidenced by a rank-order correlation of .69 between the two groups ($p < .05$). The slight differences that did exist suggest that the more highly esteemed supervisors gave somewhat more weight to industriousness and alertness while the less esteemed supervisors were inclined to look somewhat more for happiness and the individual's desire to be an efficient group member.

DISCUSSION

There is no evidence from the results of the present study to suggest that agreement between supervisors' personnel evaluations is related to the level of mutual esteem between supervisors. Furthermore, agreement between the combined supervisor evaluation and the peer-group evaluation is not significantly related to the level of esteem held for the supervisors by the peer-group. Although there was a wide range in esteem held for different supervisors (T scores of 34–70 from peers and 31–70 from supervisors to one another), the correlation values indicative of evaluation agreement were quite homogeneous statistically. The latter fact alone probably rendered any ranking of agreement levels somewhat spurious.

It is not too clear as to why agreement between supervisors is higher when the level of esteem held for both supervisors by the peer group is high. If, however, agreement between supervisors on personnel evaluations can be considered as an extension of supervisor agreement on any problem commonly facing the supervisors, the latter finding may suggest that peers tend to value supervisors who are generally in agreement with one

⁵ One supervisor was omitted from this analysis due to an incomplete set of ratings on the behavior characteristics.

another. This may be particularly true within the boundaries of an isolated and confined group where any difference of opinion on policy or problem solution can be highly noticeable to all concerned.

Finally, within the limits imposed by the particular traits and behavior characteristics used in the present study, the relatively high- and low-esteem supervisors appear to agree on the relative importance of various characteristics of personnel performance. The fact that high-esteem supervisors gave somewhat more weight than low-esteem supervisors to industriousness and alertness seems compatible with Kirchner's (1961) finding that more effective industrial supervisors stressed individual action and initiative more than less-effective supervisors while the latter were more concerned with group action and conformity. Again it is not too clear as to why the low-esteem supervisors should be in greater agreement among themselves on the relative importance of the various behavior characteristics than were the high-esteem supervisors. To the extent that each of the characteristics cited is of some value in overall performance, perhaps depending upon

the situation or the role of the person involved when displaying the characteristic, the lack of agreement among high-esteem supervisors might suggest somewhat greater flexibility on the part of such supervisors when evaluating the appropriateness or effectiveness of a particular trait or behavior characteristic.

Admittedly the present data have been derived from a small sample. One of the problems, of course, in studying unique field groups such as Antarctic stations is that small samples comprise the total population. The present problem should therefore be studied in other settings utilizing more substantial samples in order to gain further insight into the relationship between supervisor esteem and the nature of personnel evaluations.

REFERENCES

- KIRCHNER, W. K. Differences between better and less effective supervisors in appraisal of their subordinates. *Amer. Psychologist*, 1961, **16**, 432-433.
MCNEMAR, Q. *Psychological statistics*. New York: Wiley, 1955.

(Received April 1, 1963)

COMPARISON OF PROGRAMED AND CONVENTIONAL INSTRUCTION METHODS¹

MYLES H. GOLDBERG,² ROBERT I. DAWSON

Equitable Life Assurance Society, New York City

AND RICHARD S. BARRETT

New York University

The study, undertaken in an industrial setting, compared a conventional instruction method (lecture-discussion) with 2 different methods of programed instruction (programed text and teaching machine). Comparable groups of clerical employees were instructed in descriptive statistics under the 3 different instruction methods. Achievement and motivational effect were assessed at the conclusion of the course and again 6 months later. Although there was no significant difference in achievement test scores of the 3 groups upon completion of the course, the retention of information after 6 months was greatest for the group taught by the conventional instruction method. Both programed instruction methods provided a substantial saving in training time.

Much interest has been evidenced recently in programed instruction as a means of increasing the efficiency of the learning process, particularly in the light of the relative shortage of capable instructors. Recently, Hughes and McNamara (1961), Lumsdaine and Glaser (1960), and Roe (1962) have reported studies showing the effects of programed instruction in the learning situation. A number of personal experiences and viewpoints on programed instruction are expressed in articles contained in a recent American Management Association (1962) report. However, few controlled studies have been conducted in which it is possible to compare the results of programed instruction with conventional classroom instruction (lecture-discussion).

The present study compares two different methods of programed instruction (textbook and teaching machine) with each other and with the conventional classroom method in terms of both immediate and delayed recall,

time required to learn, and the trainee's attitude toward the instructional method.

METHOD

Instructional Material

The course content was the 819-frame Descriptive Statistics course, the first part of a basic program in statistics developed by Teaching Machines, Incorporated (Evans, 1960). The Programed Textbook Series format was used as a text; its machine equivalent was employed in a Ferster Tutor (Programed Teaching Aids, Incorporated—TMI) with the machine group. The conventional classroom course was designed by the instructor to cover the same course content in 18 hours (the average time reportedly needed to complete the TMI program).

Achievement Measurement

The amount of learning that took place was measured by two achievement tests designed for this study. The first achievement test consisted of 40 items developed by selecting and rewording 40 frames from TMI's Criterion Frames For Introductory Statistics. This test emphasized direct familiarity with the material as taught by the programed instruction methods.

The second achievement test consisted of 20 multipart items which primarily emphasized understanding, interpretation, and the practical application of the course content. This test was developed as follows: the instructor of the conventional classroom course constructed 60 multipart items, reflecting the course content as he would teach it. These items were carefully reviewed by the company representatives, and the 20 best items were selected. The instructor had no knowledge while

¹ All measuring instruments cited in this paper have been deposited with the American Documentation Institute. Order Document No. 7825 from ADI Auxiliary Publications Project, Photoduplication Service, Library of Congress, Washington, D. C. 20540. Remit in advance \$2.00 for microfilm or \$3.75 for photocopies and make checks payable to: Chief, Photoduplication Service, Library of Congress.

² Now at Federal Electric Corporation, Paramus, New Jersey.

teaching the course which items had been selected to make up the criterion test.

Attitude Measurement

A brief attitude questionnaire was constructed to determine in a systematic way the motivational level of trainees and their attitudes toward different aspects of their educational experience. In addition, a discussion session was held for each group where the trainees indicated how they felt about their experiences. These sessions were conducted by one of the experimenters, and two other persons acted as observers to note the comments made by the trainees.

Sample

The sample consisted of 47 beginning level clerical employees who had graduated from high school recently and were generally under 20 years of age. None had taken any prior courses in statistics, but most of their jobs involved some calculation work. Participation in this course was not completely on a voluntary basis; some individuals were assigned to take this course in much the same way as they would be assigned other work tasks.

In an effort to obtain some degree of homogeneity in the make-up of the groups, cutoff scores on the Thurstone Test of Mental Alertness (total score = 47) and the Snader General Mathematics Test (total score = 20) were used in selecting individuals for the sample. The 47 trainees were divided into three groups, well matched with respect to general mental ability and mathematical competence. Although there was no significant difference between the mean mental ability or mathematical scores of the three groups, the scores of each group covered a wide range of ability.

Procedure

Each group was scheduled for a 3-week course in statistics, meeting 2 hours a day, 3 days a week (a total of 18 hours); individuals in both programed groups were permitted to work toward the completion of their course in less than the 18 hours.

The conventional classroom course was taught by a college instructor in statistics following the content outlined in the TMI Descriptive Statistics course. No textbook was used nor homework assigned. Toward the conclusion of most sessions, the class was given a short quiz to ascertain how well they had learned the material presented in the previous session. The basic teaching approach in this course consisted of lectures, class discussions, and individual practice in problem solving.

The two programed courses (textbook and teaching machine) were identical in all respects with the exception of the format in which the material was presented. Trainees worked for 2 hours each day and at the conclusion of the work period each trainee reported the number of frames completed.

A staff member was available to handle administrative matters, but did not answer questions regarding the content of the course. No homework was assigned. Although trainees were not encouraged to take notes, note taking was permitted. If a trainee finished the program before the allotted 18 hours, he was not required to attend further meetings of his group.

After the conclusion of the course, each trainee took the two achievement tests. One week later each of the groups met separately, completed the attitude questionnaire, and discussed their feelings about the course. After a period of 6 months, the groups were retested with the same achievement tests. Due to normal employee attrition and absences, only 40 of the original sample of 47 employees were available for retest.

Efforts were made throughout the above procedure to maintain comparability in the course content. Similar explanations about the course were given to each group at their first class session; intergroup competition was avoided. Although trainees were told there would be an examination at the conclusion of the course, it was also indicated that their performance in the course would in no way affect their job situation. Despite these efforts to maintain comparability between groups, it is recognized that certain noncontrollable factors may have been operating. For instance, no time was permitted the machine group to gain familiarity with the machine operation, yet the conventional classroom approach would be familiar to all high school graduates. Perhaps more importantly, maintaining comparability of the motivational level of the three groups was particularly difficult. Also, no control was maintained over what the trainees did outside of class by way of reviewing classroom notes or studying for tests.

RESULTS

Achievement Test Comparisons

Table 1 shows the mean scores and standard deviations obtained by each of the groups on the two achievement tests. Analysis of variance of scores on the two tests showed

TABLE 1
MEAN ACHIEVEMENT TEST SCORES
UPON COMPLETION OF COURSE

Group	N	Achievement Test 1		Achievement Test 2	
		M	SD	M	SD
Classroom	14	35.8	11.7	18.9	10.9
Machine	16	41.1	8.1	16.8	8.3
Textbook	17	40.7	7.7	16.3	6.3

TABLE 2
MEAN ACHIEVEMENT TEST SCORES FOR GROUPS WHEN
DIVIDED ON THE BASIS OF THE SNADER GENERAL
MATHEMATICS TEST

Group	Achievement Test 1		Achievement Test 2	
	Top half <i>M</i>	Lower half <i>M</i>	Top half <i>M</i>	Lower half <i>M</i>
Classroom	44.6	27.0	27.9	10.0
Machine	45.5	36.1	21.6	11.4
Textbook	45.9	34.9	20.0	12.1

no significant differences in mean scores of the three groups.

To test the effect of the three methods of teaching on more homogeneous groups, each of the groups was divided in half by means of scores on the ability tests. Similar results were obtained when the groups were divided on the basis of either the Thurstone or the Snader tests. Mean scores for the top and bottom half of each group divided on the basis of the Snader test are shown in Table 2.

Analysis of variance for each of these splits showed as expected that the top half of the group (those with more mathematical ability) obtained higher mean scores on both achievement tests than those in the lower half of the group. These differences were significant at the 1% level. There were no statistically significant differences between methods of instruction.

It should be noted that the lower half of the classroom group attained a considerably lower score on Achievement Test 1 than either of the programed groups. Similarly, the top half of the classroom group scored considerably higher on Achievement Test 2 than either of the programed groups. Although these differences are not statistically significant nor can they be compared with each other in the present study, they do suggest differential effects of the methods of instruction on different ability groups, an hypothesis which can be tested in future research.

Table 3 shows the mean score results of those trainees (*N* = 40) who were given the original achievement tests upon completion of the course and who were retested about

6 months later. This table indicates that in both tests the classroom group retained more of the material than either of the programed groups. Analysis of variance indicates these differences to be significant at the 1% level.

To compare the three methods of instruction in terms of time required to complete the course, it was necessary to take achievement level into account. The mean achievement test score attained by trainees in the conventional classroom course was used as the measure of acceptable achievement for all groups. In the program groups, 76% of the trainees surpassed this level. Those trainees completed the program (machine or text) in 10–13% less time than used by the conventional classroom group (18 hours). The mean time saved by those scoring below the mean achievement score of the classroom group was 1.3 hours, or about 7% of total training time.

Attitude Comparisons

On the attitude questionnaire, trainees were asked to indicate their degree of interest in the course as of the end of each of the 3 weeks it ran. Five response categories were used: Very interested = 5, Generally interested = 4, Moderately interested = 3, Interested a little = 2, Not interested = 1. Also trainees were asked to respond to a question indicating how well they could gauge their progress during the course. Four response categories were used ranging from "all the time" to "rarely."

Table 4 shows the mean response scores of each group to these questions. The analysis of variance shows that the decreasing interest of the text and machine groups and the increasing interest of the classroom group

TABLE 3
MEAN SCORES ON ACHIEVEMENT TESTS FOR THOSE
TRAINEES RETESTED AFTER A 6-MONTH PERIOD

Group	<i>N</i>	Achievement Test 1			Achievement Test 2		
		Test	Retest	Loss (%)	Test	Retest	Loss (%)
Classroom	12	39.3	34.4	12.5	21.4	18.8	12.2
Machine	15	41.6	32.3	22.4	17.1	11.7	31.6
Textbook	13	42.2	32.8	22.3	16.6	13.3	19.9

TABLE 4

MEAN RESPONSE SCORES TO ATTITUDE QUESTIONNAIRE

Time in course	About interest in the course		
	Classroom	Machine	Text
Week 1	4.0	4.2	4.0
Week 2	4.1	3.6	3.9
Week 3	4.4	3.1	3.3
About knowledge of progress in the course			
Week 1	3.3	3.6	3.1
Week 2	2.8	3.0	2.8
Week 3	2.8	2.4	2.4

were significant ($F = 2.99, p < .05$). Differences between groups in terms of their knowledge of progress during the course were not significant.

DISCUSSION

Two different criterion measures of learning (the two achievement tests) were used in this study, and they were administered at the end of the course and again after 6 months. The advantages of programed learning were most evident at the completion of the course in helping the slower learners (as defined by lower mental ability and mathematical competence) to obtain a direct familiarity with the course material. The conventional classroom method appears to have been better in teaching understanding and the practical application of material learned in the course. It may be argued that given this objective, the program could be revised to convey this practical application as well as the classroom course. However, one implication that may be drawn, within the limitations of this specific study, is that programed methods may be most profitably used to teach course content to relatively slower learners, while the conventional classroom method may be best used to teach practical problem solving skills to relatively rapid learners.

A disadvantage of programed learning is evident in the retest 6 months later. For purpose of recall without any intervening refresher experience, the programed learning groups suffered the greatest loss. However, in a non-

experimental situation, it is most likely that employees completing a course would not wait an extended period before applying the content. Most courses in business or industry are presented to trainees so as to permit immediate application following the course.

Programed methods, because they make use of self-pacing, may result in a considerable saving in training time. Those who learned better by the programed methods did so in substantially less time. The saving in training time and in instructional costs may be the most important characteristics of the programed methods favoring their use.

Although some investigators have reported distinct advantages for either the machine or programed text methods over the other, in terms of achievement or savings in training time, no such clear-cut differences were found in this study. The machine group had the greatest degree of initial interest in the course, but this interest declined over the 3-week period. At the end of the course, the machine group showed the least interest of all three groups.

Interest in the conventional classroom method seems to have increased as the course progressed. It may be that a very important function of the instructor in the learning situation is to maintain a high level of interest in the course. Advantage may be taken of the instructor's motivational capability by providing a continuing role for him in conjunction with any programed learning course. It also seems that, although the novelty of programed method creates a high level of initial interest, the novelty and immediate knowledge of performance (often assumed to function in maintaining high interest) may not be sufficient to maintain interest throughout the program. Other motivational means, such as those provided by a capable and sensitive instructor, may be necessary for longer courses, especially where the jobs involved do not permit immediate application of newly learned material.

One of the assumed basic advantages of programed instruction is that the immediate feedback of results enables the trainees to continuously evaluate their progress in the

course. The data in this study do not reflect this advantage. The trainee taking the programmed course felt less sure of how he was doing as the course progressed. It may be important to build special examinations to be given at intervals into a programmed course so as to permit the trainee to evaluate his own progress.

Comments made in the open-ended discussion following the administration of the attitude questionnaire gave rise to two hypotheses, which may be investigated by further research. The first was that certain personality variables, such as dependency and sociability, may be closely related to which of the teaching methods is most efficient for different individuals. The second hypothesis is that increased anxiety, usually thought to be a condition which fosters learning, seems to have no effect on learning by any of the methods used in this study. The machine

group seemed to have been considerably more anxious or disturbed about the teaching experience than the other groups, and yet trainees in the machine group did not perform better than those in the other groups.

REFERENCES

- AMERICAN MANAGEMENT ASSOCIATION. *Revolution in training: Programed instruction in industry*. New York: AMA, 1962.
- EDWARDS, A. L. *Experimental design in psychological research*. New York: Rinehart, 1950.
- EVANS, J. L., GLASER, R., HOMME, L. E., & STELTER, C. J. *Descriptive statistics*. Albuquerque, N. M.: Teaching Machines, 1960.
- HUGHES, J. L., & McNAMARA, W. J. A comparative study of programed and conventional instruction in industry. *J. appl. Psychol.*, 1961, **45**, 225-231.
- LUMSDAINE, A. A., & GLASER, R. *Teaching machines and programed learning*. Washington, D. C.: National Education Association, 1960.
- ROE, A. Automated teaching methods using linear programs. *J. appl. Psychol.*, 1962, **46**, 198-201.

(Received April 1, 1963)

TRACKING WITH A DIFFERENTIAL BRIGHTNESS DISPLAY:

I. ACQUISITION AND TRANSFER ¹

STANLEY M. MOSS ²

Ohio State University

An experiment was conducted that investigated the comparative tracking performances using a differential brightness display (DBD) and a conventional positional display. The results indicated superior tracking performance with the positional display. When the percentage reduction of system error at the completion of training was computed, the differences between display groups was minimal. The inability of S to maintain precisely the same level of performance with the brightness display as with the positional display was attributed to an "area-of-uncertainty" around the reference brightness. Distinctions were made regarding performance between "early learning" and "late learning" for both displays. It was suggested that, during early learning, S established control-display relationships and learned the nature of the forcing function. During late learning, S's performance consisted of finer control adjustments. These adjustments were much finer for the positional display than for the DBD because of the area-of-uncertainty.

Fundamental to all skilled performance is the manner in which error is conveyed to the individual performing the task. The information is usually presented in the form of negative feedback in a manner that will allow the individual to make some type of corrective adjustment to minimize the system error. In most existing systems, this information is monitored by vision, although other modalities, such as audition (Forbes, 1946) and touch (Geldard, 1957; Howell & Briggs, 1959), have been investigated. In cases where the task to be performed is of a continuous nature, the error information is displayed as a point in space indicating the directionality and extent of the error. Laboratory studies of skilled performance have, for the greater part, dealt with tracking behavior. Performance error has been represented as the difference between two markers on a cathode-ray tube or the deflection of a meter needle from a reference null area. Rationale

for the use of this type of display has stemmed from attempts to represent spatial relations, such as direction, distance, and relative motion, and to optimize display-movement congruity with some operational task; that is, aircraft attitude or altitude (Fitts, 1951). The use of displays of this type has been greatly emphasized by the compatibility of their movements with control deflections. It is basically for these reasons that there has been a lack of concern for the introduction of other types of stimulus changes as a means of presenting system error. The use of new or different means of presenting error may be of interest to the theorists who are attempting to develop a theory of perceptual-motor learning. The use of positional displays may have very well limited the extent of their insight and generalizations.

In a tracking task, the ability of the subject to resolve distance and directionality of error is of basic concern in the maintenance of acceptable tracking performance. The ability of the human visual mechanism to distinguish these small spatial separations or intervals is basically a problem of visual acuity. There are other types of visual functions that could be applicable to displays of tracking tasks. The most important criterion for these variables in the design of tracking

¹ This study is in part based on a dissertation submitted to the Graduate School of the Ohio State University in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Psychology. The author is indebted to George E. Briggs for advice and assistance throughout the course of this investigation.

² Now at the Mental Health Research Institute, University of Michigan.

tasks is their ability to convey information concerning the direction and the extent of some type of stimulus change.

Brightness discrimination offers a unique method of presenting this type of information. When comparing an individual's ability for making acuity discriminations and brightness discriminations, we find that there are aspects of brightness discrimination, under certain situations, that appear to make brightness more efficient than acuity.

In the conventional tracking task, the subject has to follow the course of the target with either full head movements or eye excursions. It has been shown (Ludvigh & Miller, 1958) that there is a monotonic decrement in visual acuity as the test object moves across the visual field at increasing velocity. This decrement is explained in terms of imperfect pursuit movements of the eye which result in continued motion of the image on the retina; the continued motion results in reduced intensity contrast, a factor in producing loss in acuity. Given a task of continuous brightness change, the need to have head or eye movements is minimized. Fixation is only on that portion of the display array which shifts in brightness.

The present study was concerned with the acquisition of skilled tracking performance in which a differential brightness display (DBD) was compared with the positional-type display. In addition to the investigation of performance as a function of training, this study was concerned with the transfer effects from one display medium to the other.

METHOD

Apparatus. A conventional laboratory-tracking schema was used to investigate and evaluate the DBD. The program for this task was wired through a Donner analog computer, Model 3400. The forcing function consisted of a 12-cycles-per-minute sine wave with an amplitude of ± 7.5 volts. Control-stick movements of 7.5 inches in either direction from the null center position resulted in voltages of ± 7.5 volts. This movement corresponded to ± 15 degrees angular movement of the stick. Differences between the forcing-function voltage and the control-stick voltage were considered tracking error. Integrated absolute error was used as the primary criterion of tracking performance.

The positional display consisted of a split line generated on the face of an oscilloscope by feeding the error voltage into a time-sharing switch and then into the oscilloscope. Since the tasks were of a compensatory nature, the target of this display remained in a stationary position at the center of the vertical axis of the cathode-ray tube. The cursor fluctuated above and below the target, indicating the direction and extent of the error.

The DBD consisted of a horizontally split visual field, each half consisting of an electroluminescent (EL) lamp $1\frac{1}{2} \times 2$ inches with a $\frac{1}{8}$ -inch separation between them. The distance across the side of each of these lamps corresponded approximately to 4 degrees 46 minutes visual angle at a distance of 24 inches. These lamps were driven in the following manner:

A variable voltage bias was fed into two Airpax 400-cycle choppers. The output of one of these signals in turn was fed into one high-gain input of a Bell 20-watt stereo amplifier. The output drove the output turns of an impedance-matched output transformer. This in turn drove the lower EL lamp. The level of brightness produced by this voltage represented zero error. The error voltage from the analog computer was reduced to 14% its value and mixed with the bias voltage before it was fed into the other 400-cycle chopper. This mixed voltage followed the same course as the bias but through the second channel of the amplifier and was fed to the upper EL lamp. This produced a brightness variation above and below the brightness level of the lower EL lamp. The upper lamp was analogous to the cursor, and its brightness level was represented by the algebraic sum of the error voltage and bias voltage. The bias voltage of the lower square (target) was at the same level as the voltage needed to drive the upper lamp to a brightness that represented zero error. Thus, fluctuations in the brightness of the upper lamp which differed from the lower lamp represented tracking error: the greater the difference, the greater the error. Direction of this differential brightness (darker or lighter) represented the direction of the error. Prior to each testing session the stereo balance control was adjusted to match the brightness levels of the two lamps as requested by each subject (S).

The control stick manipulated by S moved only in the forward-backward direction. In this way control-display (C/D) compatibility was maximized for the positional display. Forcing function deflections for the positional display were represented as an excursion of 1.5 inches above and below the stationary reference. For the brightness display this was represented as a change in brightness from zero to 5.8 foot-lamberts. The reference zero error brightness was at 1 foot-lambert. Control-stick displacements of ± 15 degrees (± 7.5 inches) compensated for maximum display changes.

The displays were placed side by side on a table which had a black masonite panel with openings cut out for the displays. A sliding panel obscured one display while the other was being used. While

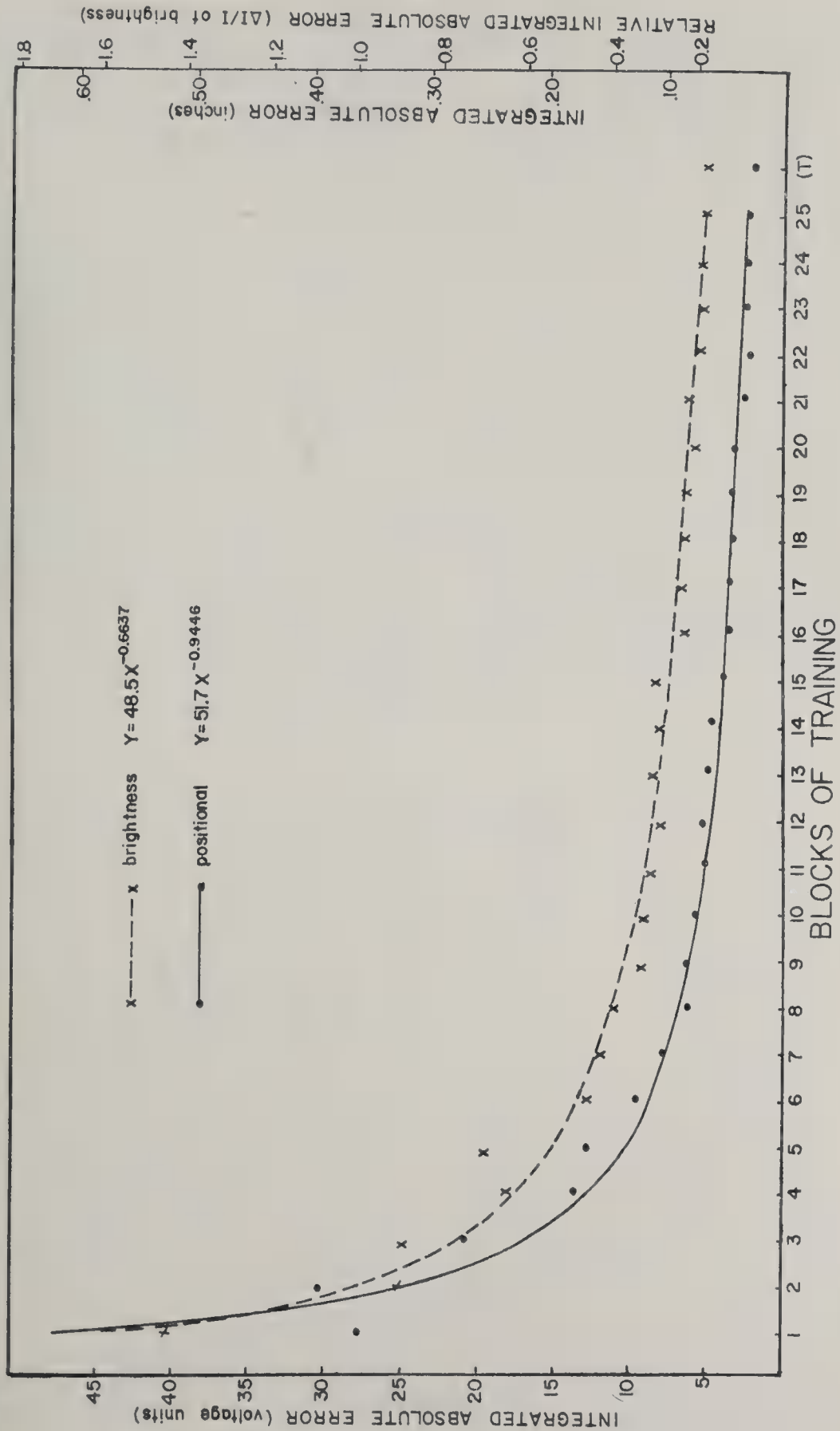


FIG. 1. Tracking performance as a function of training under both displays.

tracking, each *S* sat directly in front of his display. Ambient illumination was minimized to maximize brightness contrast for both displays.

Prior to the training and transfer conditions *S* was informed of the nature of the task and he was shown the C/D relationships for each display condition. For the positional display group the *Ss* were told that the distance between the target and the cursor represented error and that when the two lines formed a single line across the face of the cathode-ray tube they were tracking perfectly. For the brightness display group the *Ss* were told that a difference in the brightness of the upper and lower lamps represented error and that when the two lamps were of equal brightness, error was minimized.

Subjects. Twelve male undergraduate students volunteered as *Ss* for this experiment and were randomly assigned to one of two conditions: training on the differential brightness display or training on the positional display.

Daily training sessions consisted of five blocks of four 65-second trials each with a 30-second interval between trials and a 1-minute rest between blocks. Training for each display group proceeded for 5 consecutive days (25 blocks). Upon the completion of this training each group transferred over to the other display for 3 days. Transfer Days 1 and 2 consisted of five blocks each, while Day 3

consisted of four blocks on the transfer task and a final block of trials on the original training display.

Integrated absolute error scores were extracted for each *S* at the completion of each trial of training. These scores were computed during the final 60 seconds of each trial, thereby allowing *S* 5 seconds to “find” the course of the forcing function. At the completion of each trial, *S* was told his score.

RESULTS

The results of this study are summarized in Figures 1–4. Figure 1 shows tracking proficiency as a function of training with display mode as the parameter. Figure 2 shows the same information for the cross transfer with displays. The two points to the right on Figure 1 represent tracking scores on the original display after transfer to the other display. Each point in these figures is the sum of four 60-second trials for each *S*, averaged over the six *Ss* for each display group. The left-hand ordinates are in voltage units, while the two ordinates to the right are in terms of the average deviation metrics in inches and $\Delta I/I$ to

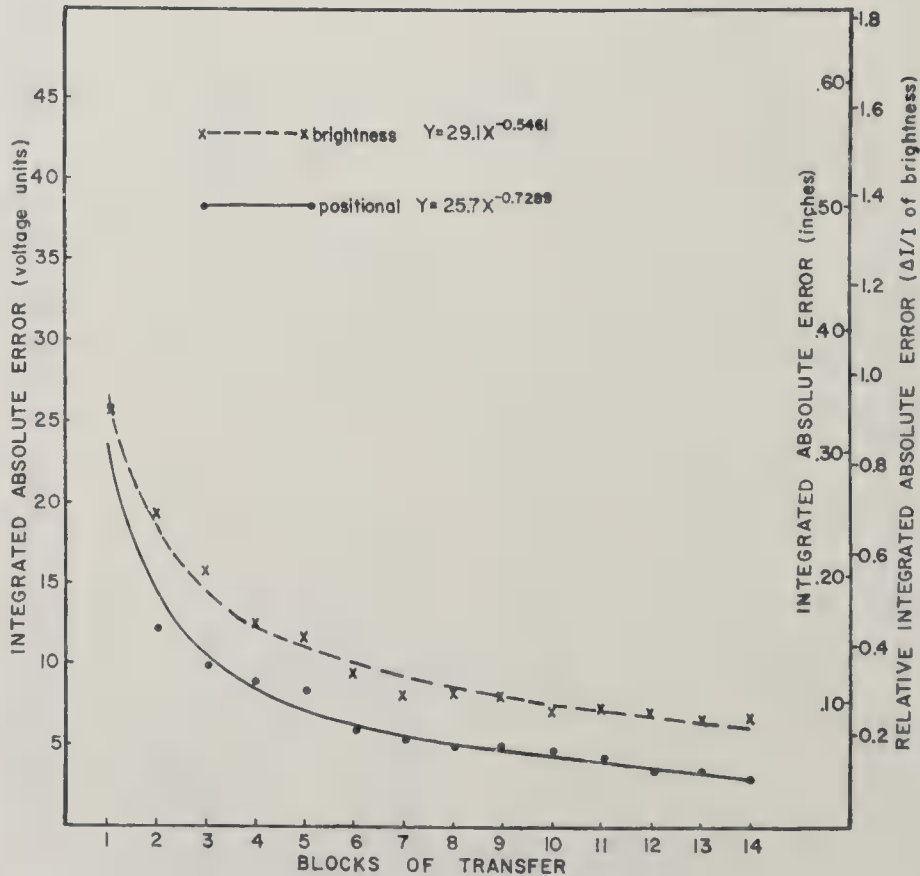


FIG. 2. Tracking performance as a function of transfer under both displays.

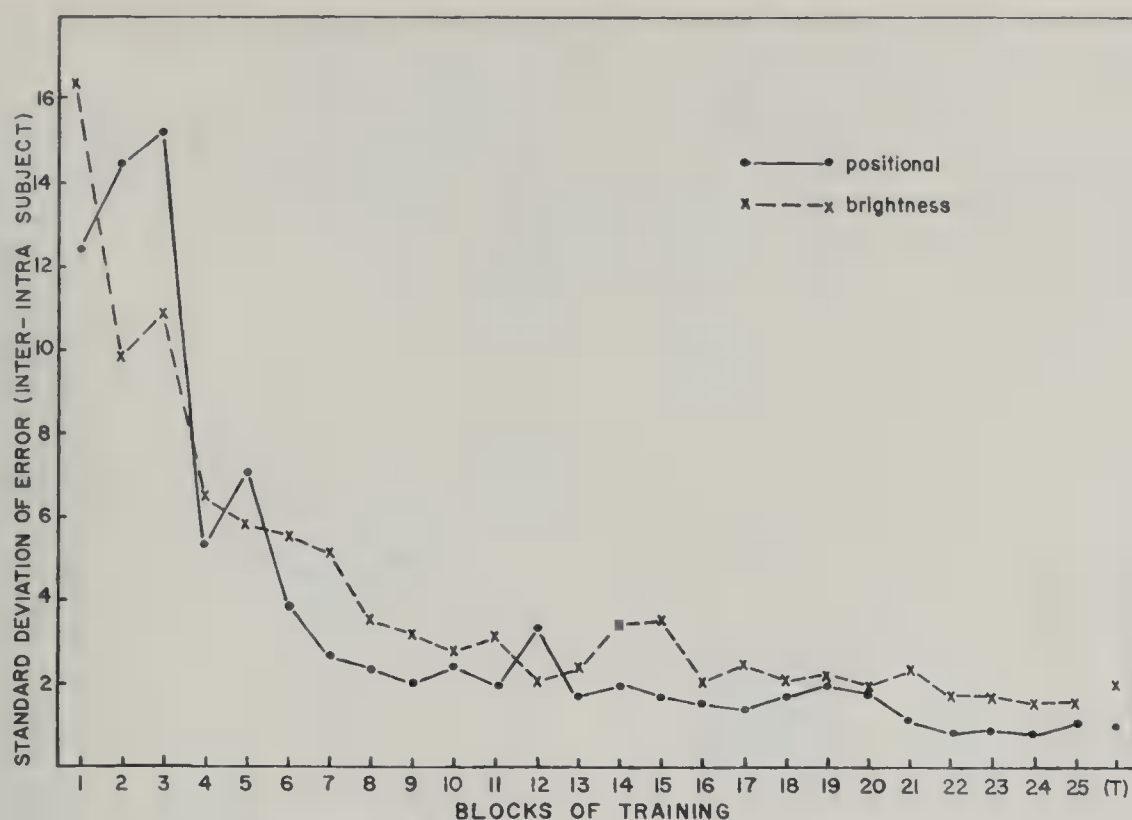


FIG. 3. Standard deviation of tracking performance as a function of training under both displays.

facilitate interpretation. The voltage units are presented to illustrate comparative performance of tracking proficiency of the two display groups, while the units to the right are presented to indicate psychophysical levels of performance for each display group. The smooth curves in Figures 1 and 2 are least-square fits for the data points for each group during training and transfer. The curves in Figure 1 are described in the functions $Y = 48.5 X^{-0.6637}$ and $Y = 51.7 X^{-0.9448}$ for the brightness display and positional display, respectively. Those in Figure 2 are described by the functions $Y = 29.1 X^{-0.5461}$ and $Y = 25.7 X^{-0.7289}$ for the brightness display and positional display, respectively. The X terms refer to the points (blocks) on the abscissa and the resultant Y terms are expressed in terms of the voltage units.

Figures 3 and 4 show the standard deviations of performance as a function of training (Figure 3) and transfer (Figure 4) with display presentation as the parameter. The two points to the right on Figure 3 represent

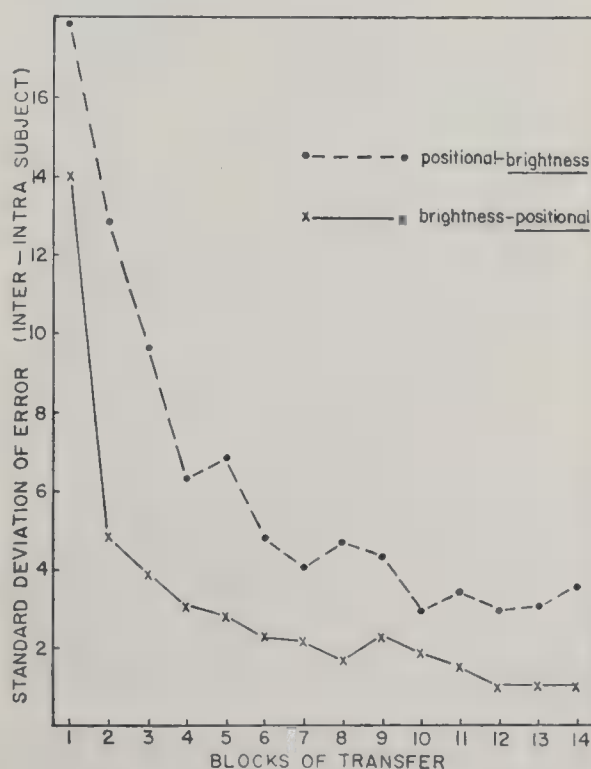


FIG. 4. Standard deviation of tracking performance as a function of transfer under both displays.

TABLE 1

ANALYSIS OF VARIANCE OF THE DATA FOR THE LAST FIVE BLOCKS OF FIGURE 1—TRAINING

Source	<i>df</i>	<i>MS</i>	<i>F</i>
Between Ss	11	24.97	
Displays(D)	1	207.58	30.94**
Error	10	6.71	
Within Ss	48	.48	
Blocks(B)	4	1.45	3.92*
B × D	4	.57	1.45
Error	40	.37	

* *p* < .05.
** *p* < .01.

these values for the original display groups after transfer. Each point in these figures is the combined intra-inter-*S* standard deviation calculated from the four 60-second trials of each *S* across the six *Ss* (*N* = 24).

Separate analyses of variance were applied to the last five blocks of the training session and to the last four blocks of the transfer session. The values used in computing these analyses were *S* means for each block of training and transfer. The analysis of the training data (Table 1) shows that the main effects of Displays and Blocks are significant at *p* < .01. The analysis of the transfer data showed that the main effect of Displays was significant at *p* < .05.

An analysis of variance also was applied to the data between the last Block of the original learning and the first Block of the transfer sessions. The main effects consisted of Groups and Sessions. The two groups consisted of: Group I, those *Ss* who trained originally with the positional display and then transferred over to the brightness display; Group II, those *Ss* who trained originally with the brightness display, and then transferred over to the positional display. The two sessions consisted of the original learning and the transfer condition. The Groups × Sessions interaction then would indicate any cross-display transfer effects. Only Sessions attained significance, *p* < .001.

A *t* test was applied to the difference in performance scores between Block 25 of the original training sessions and retest on the original display (Block 15 of the transfer

session). The resulting *t* was not significant (*t* = .471, *df* = 8).

Upon inspection of Figures 1 and 2 it is apparent that performance, as described in voltage units, with the positional display is superior to performance with the brightness display for the original training and for the transfer session. After the first several blocks under each condition, improvement in performance as a function of practice appears to be comparable for both display groups. This is illustrated by fitting a linear equation to the last 20 blocks of the original learning and the last 9 blocks of the transfer condition for each display group. The application of this equation is used only as a means of demonstrating similarities in the slopes of the display groups after the first few blocks of training. The slope constants for the original learning are −.39 (*p* < .01) for the brightness group and −.35 (*p* < .01) for the positional group. Those for the transfer condition are −.33 (*p* < .01) for the brightness group and −.37 (*p* < .01) for the positional group. These values differ considerably from the slope constants generated by the best-fit equations. It is interesting to note that the differences between the slope constants of the two display groups for the best-fit equations are considerably larger than the differences between the slope constants of the two display groups for the linear equations. For this reason it is felt that a distinction can be made between “early learning” and “late learning.”

This is given support by a closer look at Figure 3. The standard deviation of performance scores for Blocks 1–5 shows considerable overlap between the display groups. As training progresses beyond this point, the differences between the groups become more distinct, with the positional display group reflecting less variability in performance. Explanations for these differences lie in the fact that during early learning *S* establishes C/D relationships, and learns the nature of the forcing function. Late learning would then consist of finer-control adjustments as reflected by the continuing improvement in performance. Display-group differences during early learning can then be attributed,

for the greater part, to the greater difficulty in establishing C/D relationships and extracting the true nature of the forcing function for the differential brightness group. This difficulty persists beyond early learning as indicated by the slightly larger value of the slope constant for this display during late learning. Comments by *S* lend support to this.

DISCUSSION

Fleishman and Hempel (1953) present evidence clarifying the distinction between "early" and "late" performance on a perceptual-motor task. These investigators factor analyzed a criterion motor task at different stages of practice. They found a systematic change in the factor structure as practice continued. During early stages of practice nonmotor factors contributed most of the variance, while motor factors contributed practically all of the variance during late stages of practice. These results are in agreement with speculations drawn from the present study; early learning consists of establishing C/D relationships and learning the nature of the forcing function (non-motor), and late learning the establishment of finer control movements (motor).

This inability of *S* to exercise more accurate control with the brightness display may introduce another factor which would expand the differences in performance between the display groups. The momentary loss of the forcing function would induce a rather rapid change in the brightness level of the display, either above or below the reference brightness. In the former case this rapid increment in brightness would change the adaptation level of the already display-adapted eye. Hence, immediately subsequent matching would not be exact because of the decrease in sensitivity of that part of the retina that fixated on the variable portion of the display.

By the time *S* completed the original training session he had attained a relatively high level of skill in tracking even though there are differences between the display groups. Upon transfer to the other display, *S*'s early-learning performance did not exhibit the

same variations as it did in the original learning. Figure 4 shows no overlapping in the variability of performance between the display groups, the positional display group indicating less variability throughout the entire transfer condition. It appears that C/D relations for the positional display are attained very early, while C/D relations for the brightness group are not. Again, *S* reported difficulty in tracking with the brightness display.

It is interesting to note the levels of performance attained at the completion of the transfer (Block 14, Figures 2 and 4) as compared with performance during original learning (Blocks 14–17, Figures 1 and 3). On the one hand, performance with the positional display at the completion of transfer is comparable to mean performance at Block 17 with the positional display during original learning, while variability of performance at the completion of transfer is comparable to the variability of performance at the completion of the original learning. On the other hand, mean performance with the brightness display at the completion of transfer is comparable to Block 17 of the original learning, while variability in performance is comparable to Block 15 of the original learning. The *Ss* who performed with the brightness display during transfer were functioning at about the same level of proficiency as the original group achieved during original learning for that display condition. The amount of training was the same for each group. The *Ss* performing with the positional display during transfer exhibited the same mean performance level but with a lower degree of variability for comparable amounts of training.

It should be noted that at the completion of this study several *Ss* reported difficulty finding the "center" of the reference brightness, while no statements were made regarding "lining up" the cursor with the target in the positional display. For this reason, it can be assumed that the subjective tolerances around zero error were much greater for the brightness display. Following this, it would be reasonable to assume during the training portion of this study that the *Ss* tracking

with the positional display were making finer-control adjustments to keep the cursor in coincidence with the target. Those Ss tracking with the brightness display during this period were making smoother control movements because of an area of uncertainty around zero error. Transferring over to the positional display, these Ss would, once C/D relationships were established, immediately make the fine control adjustments in lining up the cursor with the target. Those Ss transferring over to the brightness display, once C/D relationships were established, would persist in making fine control adjustments in an attempt to find the center of the reference brightness—this persistence being due to the “higher” scores reported to them by the experimenter at the completion of each trial with the brightness display, as compared with their performance scores reported to them while training with positional display. In attempting to narrow this area of uncertainty by fine adjustments, rapid variations may be induced causing loss in retinal sensitivity, as stated earlier, and hence greater variability of performance.

Although the differences in performance between the display groups exist, the size of the system error with either display is small. When the percentage reduction of system error at the completion of training was com-

puted, the difference between display groups was only 5%. The significance of this difference depends on the overall performance criteria established for the system. For some control systems this difference may be negligible. On the other hand, it should be kept in mind that performance was evaluated using a coherent (simple sine wave) input which, when taken out of the laboratory, presents an unrealistic situation.

REFERENCES

- FITTS, P. M. Engineering psychology and equipment design. In S. S. Stevens (Ed.), *Handbook of experimental psychology*. New York: Wiley, 1951.
- FLEISHMAN, E. A., & HEMPEL, W. E. Changes in factor structure of a complex psychomotor test as a function of practice. *USAF Hum. Resour. Res. Cent. res. Bull.*, 1953, No. 53-68.
- FORBES, T. W. Auditory signals for instrument flying. *J. aero. Sci.*, 1946, 13, 255-258.
- GELDARD, F. A. Adventures in tactile literacy. *Amer. Psychologist*, 1957, 12, 115-124.
- HOWELL, W. C., & BRIGGS, G. E. An initial evaluation of a vibrotactile display in complex control tasks. *Ohio State U. Res. Found. tech. Rep.*, 1959(Oct.), No. (813)-5.
- LUDVIGH, E., & MILLER, J. W. Study of visual acuity during ocular pursuit of moving test objects: I. Introduction. *J. Opt. Soc. Amer.*, 1958, 48, 803-808.

(Received April 1, 1963)

CHARACTERISTICS OF SUCCESSFUL POLICEMEN AND FIREMEN APPLICANTS¹

JOSEPH D. MATARAZZO, BERNADENE V. ALLEN, GEORGE SASLOW,
AND ARTHUR N. WIENS

University of Oregon Medical School

The Ss were 243 successful applicants for the positions of policemen and firemen who were placed on the eligibility lists for these 2 positions during 1959-62. Each was studied individually in an 8-hr. examination which included the WAIS, MMPI, EPPS, SVIB, Rorschach, 5 other instruments, and an interview. The results reveal that these successful Civil Service applicants are of high intelligence (IQ 113) and have superior personality adjustment. Interests of policemen applicants are social service in orientation, while those of firemen are primarily technical and business in orientation. The study reveals that, at least in 1 large city, fire and police departments are recruiting superior young men into their ranks.

Although there appears to be a national interest in selecting more qualified policemen and firemen (Frost, 1955; Oglesby, 1957), relatively few studies have been reported of applicants for, or persons already in, these two occupational groups. According to Chenoweth (1961, p. 232) the policeman selection procedure utilized almost universally today is very little different from the one first employed for police selection in London, England, in the year 1829, which was a character check, medical examination (including some estimate of the applicant's intelligence), and a personal interview. However, a 1955-56 mail questionnaire survey of chiefs of police of 90 United States municipalities (an 81% return of those sampled) with a population of 100,000 or more, indicated some changes in the past decade, since it revealed that 14 cities (16%) were using some type of psychiatric or psychological screening of police applicants (Oglesby, 1957). The author conducted personal interviews of the key officials in each of these 14 cities and found that Wilmington and

Toledo appear to share the honor of having the oldest continuous United States program of psychiatric screening of police applicants. Both of these began in 1938; followed, in the next 18 years, by 12 other cities. Of the 14 cities, 10 employed in screening only a psychiatrist, 2 only a psychologist, and 1 both a psychiatrist and a psychologist (the author did not report similar data for the fourteenth city).

Although not reported by Oglesby, as is apparent from several studies published during the past decade, some cities (notably St. Louis and Baltimore among the earliest ones) have begun to include modern methods of objective and projective psychological assessment presumably carried out by psychologists. These studies reveal the use in psychological assessment of such additional procedures as (a) the Army General Classification Test (AGCT; DuBois & Watson, 1950; Dudycha, 1955, p. 59; Mullineaux, 1955); (b) a specially devised police aptitude test, a short open-ended composition test "Why I Wish to Become a Policeman," and the Cornell Word Form (DuBois & Watson, 1950); (c) the Rorschach test and the Strong Vocational Interest Blank (SVIB; Kates, 1950); and (d) the Kuder Preference Record (Sterne, 1960). Personal correspondence indicates that an increasing number of cities are using psychological assessment procedures, even though few reports of these

¹ We wish to acknowledge the complete freedom to carry out this study and numerous other forms of invaluable help provided by John Montague, MD; the heads of Portland's Civil Service Board; Police and Fire Bureaus; Board of Trustees of the Fire and Police Disability and Retirement Fund; and officers of Portland's Local No. 43 of the International Firefighter's Association. Gerald Solomon was of invaluable aid in the analysis of the data.

have as yet been published. While the majority of cities utilizing a psychological (or psychiatric) interview of applicants presumably utilize an interviewer who conducts the interview with the applicant alone, those cities which also employ standardized objective and projective techniques of assessment (AGCT, SVIB, Kuder Preference Record, Rorschach test, et al.) appear to utilize group administration of these selection devices. Such group examination procedures obviously introduce economies both in the amount of professional time required for assessment and the cost of the service. However, group assessment procedures do introduce many problems which limit their effectiveness for other than very crude selection decisions.

Assessment of firemen applicants appears to have received only scant attention as evidenced by published studies (Anikeef & Bryan, 1958). However, some data on firemen and policemen, collected incidental to other purposes, have been published (Stewart, 1947, pp. 13-14; Thorndike & Hagen, 1959, pp. 241-246).

In 1959 the city of Portland (Oregon), concerned by the very real personal, as well as administrative, costs of an occasional psychiatric breakdown in one of its policemen or firemen, instituted an intensive program

of psychological assessment. This was carried out by the Department of Medical Psychology at the University of Oregon Medical School and was integrated into a large research program currently underway. The present paper is a report of the characteristics of the 243 applicants evaluated by us in the 3-year period, 1959-62.

All of these men had passed a prior Civil Service written examination, medical examination, and a departmental and/or Civil Service oral interview. Before the introduction of the psychological assessment program in 1959, as vacancies occurred in the two departments, individuals from this successful sample of 243 would have been appointed to the police or fire departments, since all other applicants had failed one or another of the prior screening procedures (Civil Service, medical, or oral). Thus, a study of the characteristics of these 243 successful applicants provides a good index of the characteristics of men who have been appointed policemen and firemen in the city of Portland from 1959 to 1962 and who, presumably, have some characteristics in common with successful applicants for these positions in other large cities which use only the standard Civil Service, medical, and oral examination procedures.

Table 1 provides a summary of the prior

TABLE 1
DERIVATION OF THE 243-MAN SAMPLE STUDIED

Position	1959	1960	1961	Total
Policemen				
Applications received by Civil Service	336	554	1038	1928
Applications approved by Civil Service	270	540	1012	1822
Applicants passing Civil Service written examination	93	130	261	484
Applicants passing physical agility test	82	75	150	307
Applicants passing police department interview	65	51	86	151
Applicants passing medical examination	28	42	64	133
Applicants referred for psychological assessment	28	42	46	116
Firemen				
Applications received by Civil Service	560	618	710	1888
Applications approved by Civil Service	459	588	662	1709
Applicants passing Civil Service written examination	149	155	203	507
Applicants passing physical agility test	100	104	133	337
Applicants passing medical examination	51	83	126	260
Applicants referred for psychological assessment	24	35	68	127

screening history of this sample of 243 successful applicants who finally were placed on the eligible lists for the two positions. The sample of 243 includes 116 successful policemen and 127 successful firemen applicants. The rates of attrition during the screening process shown in Table 1 are comparable to those reported for policemen by DuBois and Watson (1950) for the city of St. Louis.

The present paper is designed to provide information on the personal and psychological examination characteristics of these 243 men. With analyses now underway, we hope in the future to identify the characteristics of the men from this group of 243 who, after 1959, were passed by us following an 8-hour individually administered clinical-psychological-assessment procedure (56% of the 116 heretofore successful policemen candidates and 57% of the 127 heretofore successful firemen candidates) and compare these characteristics with those of the 44% and 43% samples, respectively, of men whom we rejected after this intensive clinical psychological study.

METHOD

Subjects

To be eligible for either of the two Civil Service positions, the 243 successful applicants had to be between 21 and 29 years of age (upper limit to age 34 for veterans) and have completed high school or have the General Educational Development (GED) Test equivalence of such education. As noted earlier, each of the 243 applicants had taken Civil Service written examinations (including the Wonderlic) as part of the prior screening procedure. The applicant is typically not rejected for a low score on this or any of the other routine screening procedures (except for the medical examination). Instead, low-scoring men are placed in appropriately higher-numbered positions on the list of 243 eligible men. The best-scoring man is given the rank of 1 and is placed first on the eligible list.

Procedure

The 243 successful candidates were sent to the University of Oregon Medical School, for intensive psychological assessment, lasting 7-8 hours per candidate. Each of the 243 men was evaluated individually and, for all but one procedure per candidate (the interview), by the same clinical psychologist. For each of the 243 men, the standardized individual examination consisted of the following: Wechsler Adult Intelligence Scale (WAIS),

Rorschach Inkblot Test, Miale-Holsopple Sentence Completion Test, Taylor Manifest Anxiety scale, Saslow Psychosomatic Inventory, Cornell Medical Index, California F Scale and Adorno Authoritarian F Scale, SVIB ($N=236$), and the Edwards Personal Preference Schedule (EPPS; $N=172$). The N s did not add up to 243 on each of the last two inventories because one or another of them was omitted at the end of the all-day session in cases where the man had arrived late, had to leave for his late-afternoon job, had been unusually slow, etc. From 1961 on, the Minnesota Multiphasic Personality Inventory (MMPI; $N=84$) has been substituted for the EPPS, and, since 1962, we also have added the Maudsley Personality Inventory.

At the beginning of the all-day assessment, each candidate filled out a 1-page biographical sketch listing his pertinent familial, educational, and occupational history. This information was utilized later in the day by one of two interviewers, a psychologist (J.D.M.) or a psychiatrist (G.S.) who, after a study of the considerable assessment data obtained by the clinical psychologist, and a conference around this data by the three of them and often one other clinical psychologist (A.N.W.), would conduct a 45-minute clinical interview of the applicant. The interview was designed to provide an intensive psychiatric examination of each candidate's current psychological make-up and, especially, to identify any signs of current or potential emotional instability. Despite our recognition of the many shortcomings of the typical psychiatric interview (see Matarazzo, 1964, for a review of this subject), we felt that, in conjunction with all of the other assessment data on each man available to the interviewer, such an interview was invaluable for integrating all of this other data into one meaningful whole for each applicant. At the end of the day, the three or four clinicians independently would rate each man as a "pass" or "fail." Only the men receiving a pass rating have been retained on the two departmental eligibility lists since 1959. Although the data are still being analyzed and will not be reported here, the reasons for failure (i.e., candidate tagged as falling in a "high risk" group) for any given individual could have been revealed in any 1 or more of the 10 objective or projective examination procedures, or from the interview. Almost without exception, individuals placed in the "probability of a high risk," or failure, group were so designated because of negative indicators in both the interview (emotional instability, immaturity, psychopathic personality, psychosis, etc.) and a number of the other procedures (e.g., bizarre responses on the Rorschach, unusually high scores on the anxiety and psychosomatic inventories, clinically fragile MMPI profiles, etc.). Only with further research will we be satisfied that we have demonstrated acceptable validity correlates for our up-to-now clinical method of selecting these low risk (pass) and high risk (fail) groups from the 243 applicants described in the present study.

TABLE 2
CHARACTERISTICS OF CIVIL SERVICE AND DEPARTMENT-APPROVED POLICEMAN
AND FIREMAN APPLICANTS FOR 3 SUCCESSIVE YEARS

		Means									
		Position		1959		1960		1961		Total	
N ^a		28	24	42	35	46	68	116	127		
Age	Police	26.2		25.6		25.5		25.7			
	Firemen	24.6		25.3		24.3		24.6			
Education	Police	12.9		12.7		12.5		12.6			
	Firemen	12.7		12.9		12.8		12.8			
WAIS Verbal scale (IQ)	Police	110.2		114.0		111.8		112.2			
	Firemen	109.3		112.3		112.9		112.0			
WAIS Performance scale (IQ)	Police	107.7		111.2		115.0		111.9			
	Firemen	111.8		111.5		112.3		112.0			
WAIS Full scale (IQ)	Police	109.8		113.6		114.1		112.9			
	Firemen	111.0		112.8		113.4		112.8			
Taylor Manifest Anxiety scale	Police	7.1		6.0		5.3		6.0			
	Firemen	5.8		6.3		6.5		6.3			
Saslow Screening Inventory	Police	2.9		2.8		2.9		2.8			
	Firemen	3.1		3.1		2.9		3.0			
Cornell Medical Index	Police	6.6		5.3		4.8		5.4			
	Firemen	3.3		4.8		4.8		4.5			

^a First number of each pair is policemen; second number is firemen.

RESULTS

Age and Education

As is shown in Table 2, the mean age of the 116 policemen and 127 firemen applicants was 25.7 and 24.6 years, respectively. (Interestingly, this small difference, significant at the .01 level, is the only such statistically significant difference in the combined total sample shown in Table 2.) The ages of the 243 applicants covered the total possible range (21–29) and included 22 applicants (9%) in the up-to-age-34 veteran-eligible group (15% policemen and 4% firemen in this older group).

Educational level did not differentiate successful policemen and firemen applicants (means of 12.6 and 12.8 years, respectively). In view of the fact that the means for both groups indicate almost a year of college for the average applicant for these two positions, the actual educational attainment by year completed is of interest. For the 243 men this was as follows: 9 years of school completed, 2%; 10, 7%; 11, 5%; 12, 46%;

12–13, 6%; 13, 12%; 13–14, 3%; 14, 10%; 14–15, 1%; 15, 5%; and 16, 3%. (Policemen and firemen groups yielded almost identical percentages in each educational level.) Thus, these results show that 14% of the 243 men had not finished high school (they substituted the GED Test equivalence for the high school diploma) and 40% had had some college education (with 3% of the total of 243 having earned a college degree). Clearly, these results indicate that the applicants for these policeman and fireman positions are neither dull nor uneducated.

Intelligence

In view of the surprising percentage of college-educated applicants for these two positions (40%), the findings for intelligence shown in Table 2 are easier to understand. Thus, with a WAIS Full Scale IQ of 112, the average policeman and fireman applicant in the city of Portland who finally makes the eligibility list falls at the eightieth percentile; that is, he is more able intel-

lectually than 80 out of every 100 men and women in the general population. This figure is approximately the IQ level of the average college graduate in the United States.

We feel that this finding is of such importance that the full range of IQ in our 243-man sample should be reported. This is done in Figure 1, where each square represents the Full Scale IQ of one of the 243 applicants. Figure 1 further reveals that the IQ scores ranged from 85 to 130, with a median of 113 (in contrast to the mean of 112). Thus, all but 4 (1.6%) of the 243 applicants had WAIS IQs of 100 or above, in contrast to the 50% of such scores expected in the general population.

Clearly, then, a city such as Portland, even without the help of modern psychological assessment procedures, currently is appointing to its policemen and firemen ranks men of considerably above average intelligence.

Emotional Adjustment

The results with three paper-and-pencil inventories also are shown in Table 2. Norms for the Taylor Manifest Anxiety Scale, Saslow Psychosomatic Screening Inventory, and the Cornell Medical Index collected by us on normal and patient groups (Matarazzo, Matarazzo, & Saslow, 1961, p. 57), as well as norms published by the individual authors of these scales, indicate that the findings shown in Table 2 place our 243 young men at the very healthy end of each scale. In fact, in terms of self-reported emotional and personality adjustment these 243 Civil Service applicants report fewer symptoms than do medical students of the same age from our own school's graduating class of 1961 and 1962 (Kole, 1962).

That the 243 Civil Service applicants should score so well on emotional ad-

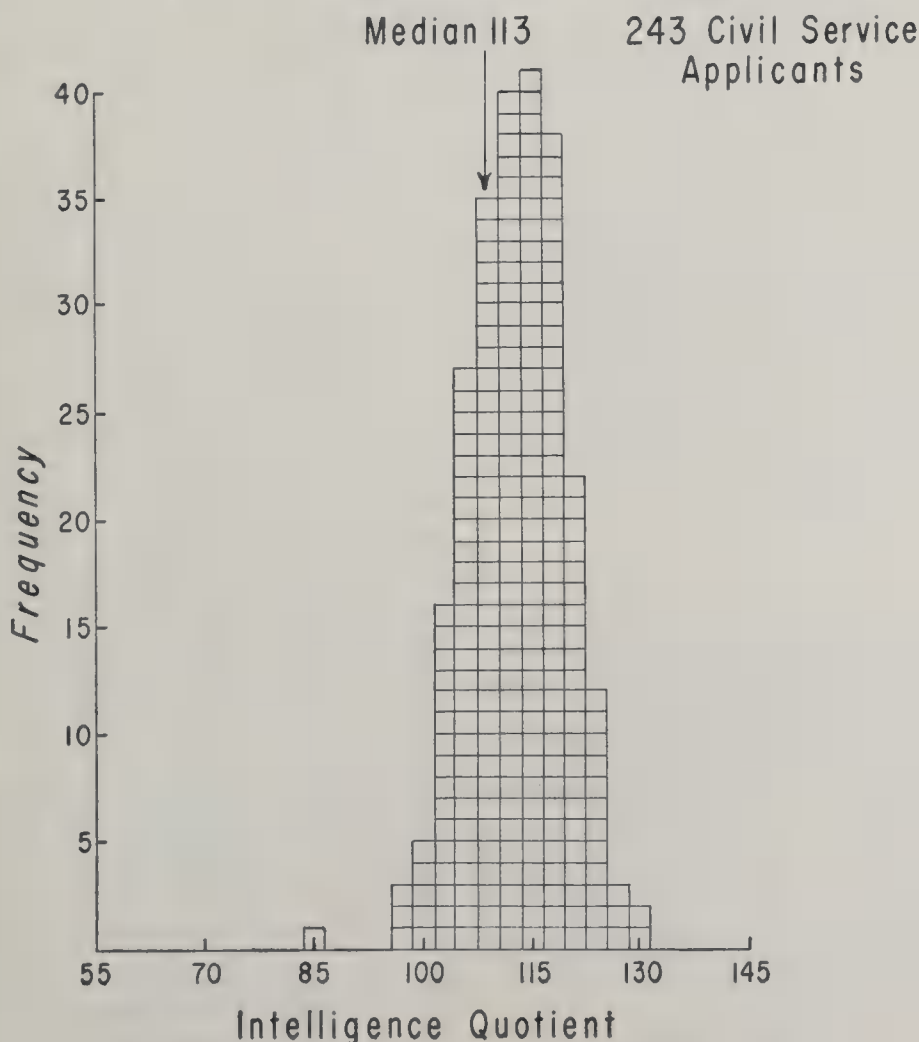


FIG. 1. IQ: Full WAIS for 243 Civil Service applicants.

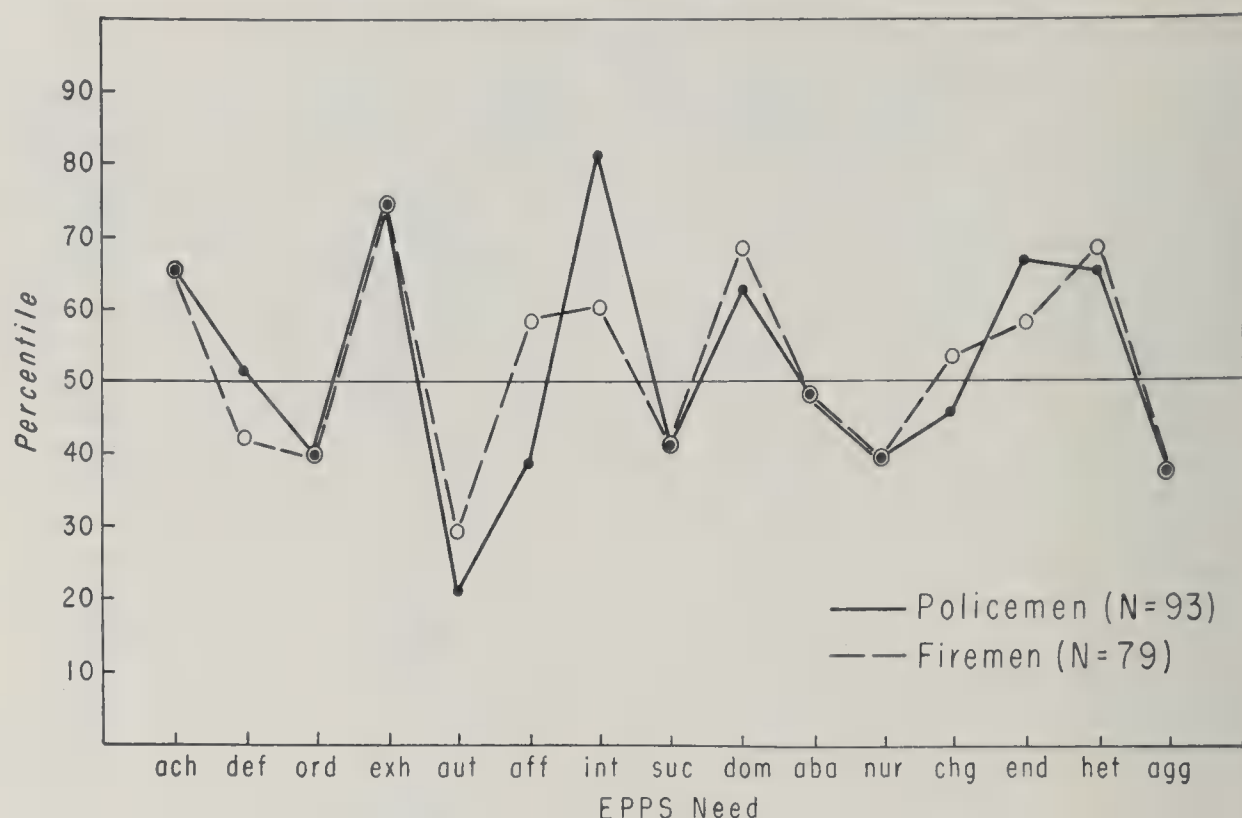


FIG. 2. EPPS profiles for policeman and fireman applicants.

justment is probably a reflection of (a) their wish and ability to score well on such self-report, personal adjustment, job-selection inventories; and (b) the important fact that each of these men had passed a prior medical examination as part of the selection process. In any event, measured by what they report about themselves on such inventories, the average policeman and fireman applicant describes himself as better adjusted than the average person his age.

Personality Needs

These were assessed by the EPPS, an inventory designed to tap 15 needs which Edwards (1959) and others believe represent some of the most important of the human needs involved in everyday personality functioning. Profile results from this inventory are shown in Figure 2. The profiles shown are expressed in percentile values using, as a reference group, a sample of 4,031 men from a nationwide sample of men from the general adult population (Edwards, 1959, p. 13).

The profiles in Figure 2 show that (a)

relative to each other, and for all practical purposes, the firemen and policemen groups of applicants do not differ materially in these 15 needs; and (b) relative to men in general, both groups of applicants are higher than the average man in their needs for Achievement, Exhibition, Intracception (ability to analyze and understand the feelings and behavior of others), Dominance, Endurance, and Heterosexuality (masculine interests) and lower than average in Autonomy (need to work independently), Succorance (need for encouragement, kindness, and help from others), Nurturance (need to forgive, sympathize with, or to help friends and strangers who are sick or in trouble), and Aggression (need to criticize others, or tell them off, or get revenge).

Thus, both groups of applicants describe themselves as having strong needs to excel or achieve, be the center of attention, understand and dominate others, stick to a job until it is done, and to "be one of the boys" among men. Along with this, their lower than average needs suggest that both groups of men like to work with others rather than

autonomously, need little kindness or succorance from others, in return give little sympathy to others, all the while feeling little animosity or aggression toward their fellow man.

Interestingly, and not surprisingly since they work and live together in fire stations, Figure 2 shows that firemen applicants, relative to police applicants, have a higher than average need for Affiliation (the need to be and work with others). They also have a slightly higher need for Change (to do new and different things, participate in new fads, etc.) than do policemen applicants. Policemen applicants, on the other hand, have a slightly higher need for Intraception (ability to analyze and understand the feelings and behavior of others).

Interests

The results with the SVIB (Strong, 1943) are shown in Figure 3. Occupational scales are grouped in Figure 3 according to the method described by Hagenah (1960). Both standard scores and letter grades are presented in order to facilitate interpretation of the findings.

The results in Figure 3 show that: relative to firemen applicants, the interests of policemen applicants are more like those of (a) persons already employed in the Social Service occupations (Group V), with 8 out of the 8 differences in the Group V occupations reaching the .001 level of confidence; (b) Psychologists (.01 level); (c) Policemen (.001 level); (d) Senior CPA (.05 level); and (e) CPA (.01 level); whereas firemen applicants exceed policemen applicants on the occupational scales entitled Engineer (.01 level); Farmer (.05 level); Industrial Arts Teacher (.05 level); Real Estate Salesman (.01 level); and President, Manufacturing Concern (.01 level). The two groups of Civil Service applicants do not differ on the remaining 30 of the 48 scales in the Strong.

Thus, policemen and firemen applicants, while having interest patterns with considerable overlap, do reveal some differences which suggest that young policemen are more oriented toward jobs involving working with people (Group V), while young firemen are

oriented towards occupations requiring work with one's hands or the business world. The clinical interviews with these applicants bore out these findings since, typically, the policeman applicant stated that his reasons for choosing this line of work included his wish to work with juveniles, or with men on probation, etc., while firemen applicants often stated they chose firefighting because the 24-hour-on, 48-hour-off work schedule permitted them to farm, hunt, fish, buy, renovate, and sell old homes for profit, work as real estate salesmen on their days off, etc.

It was our clinical impression that firemen and policemen applicants do differ, with the former being the rugged, outdoor, family handyman type of person and the latter the more intellectual, professional-type person. The differences between the two groups obtained with the SVIB, more than with any other assessment technique used by us, most clearly reflected the differences we noted clinically. (Further evidence, that the self-concept of young police applicants is like that of, for example, the white collar probation officer while that of the young firefighter is that of the rugged outdoorsman, is that we soon came to recognize representatives of each applicant group by their dress.) In a sample of 106 applicants, policemen more often came to the examination dressed in suit, white shirt, and tie (52%) whereas firemen did so much less often (15%). Conversely, sport-shirt wearers in the same two groups were 48% and 85%, respectively ($\chi^2 = 15.84$, $p = .001$).

Personality Profile

While we have continued to use the Rorschach Inkblot Test for clinical evaluation of these applicants, the results with this assessment technique will not be reported here other than to point out that, using Beck's method of scoring each protocol, the group of 116 policemen and 127 firemen applicants did not differ on any Rorschach variable.

As mentioned earlier, during 1961 we substituted the clinically more potentially useful MMPI for the EPPS. The mean MMPI findings for 84 applicants are shown in

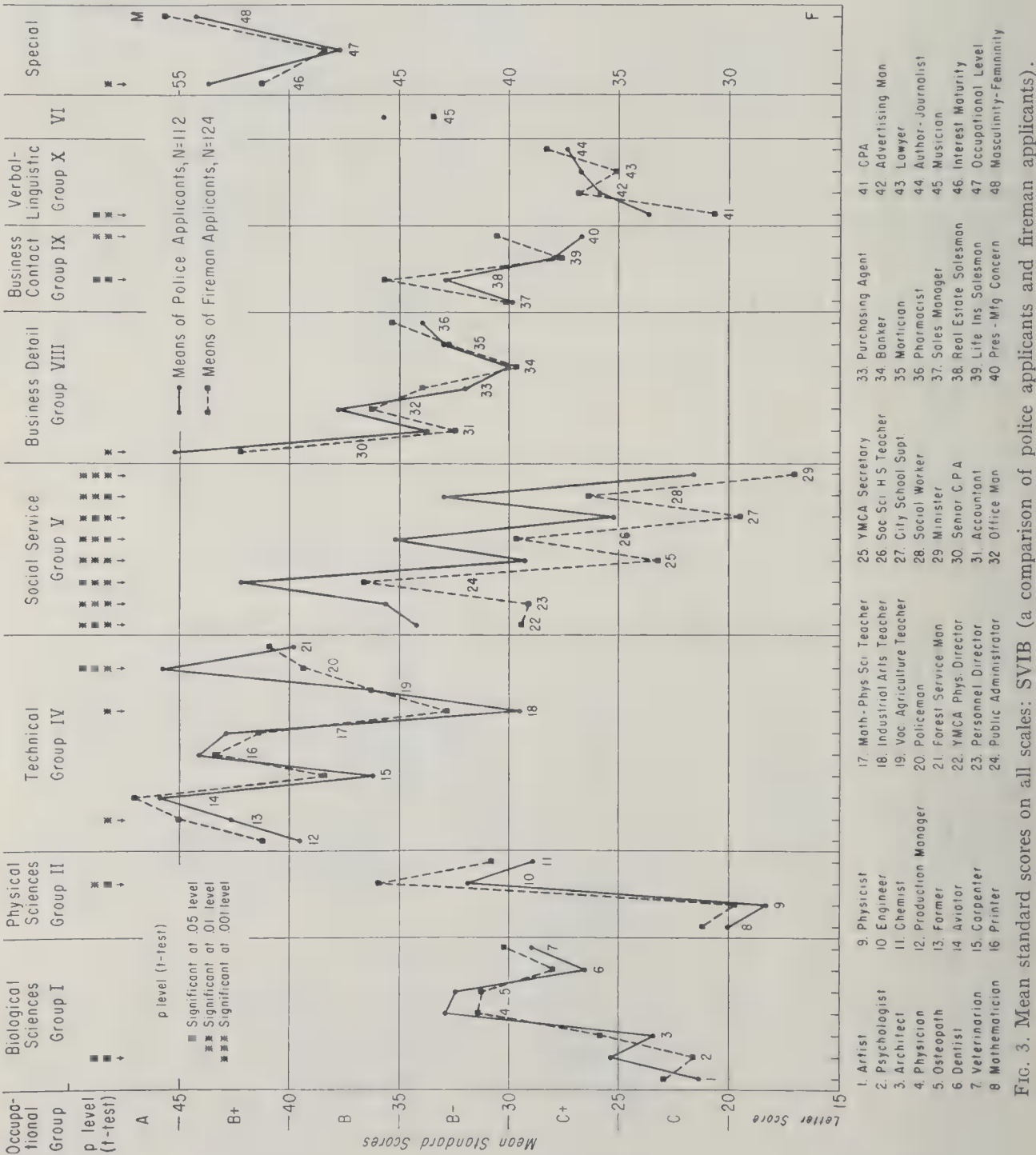


FIG. 3. Mean standard scores on all scales: SVIB (a comparison of police applicants and fireman applicants).

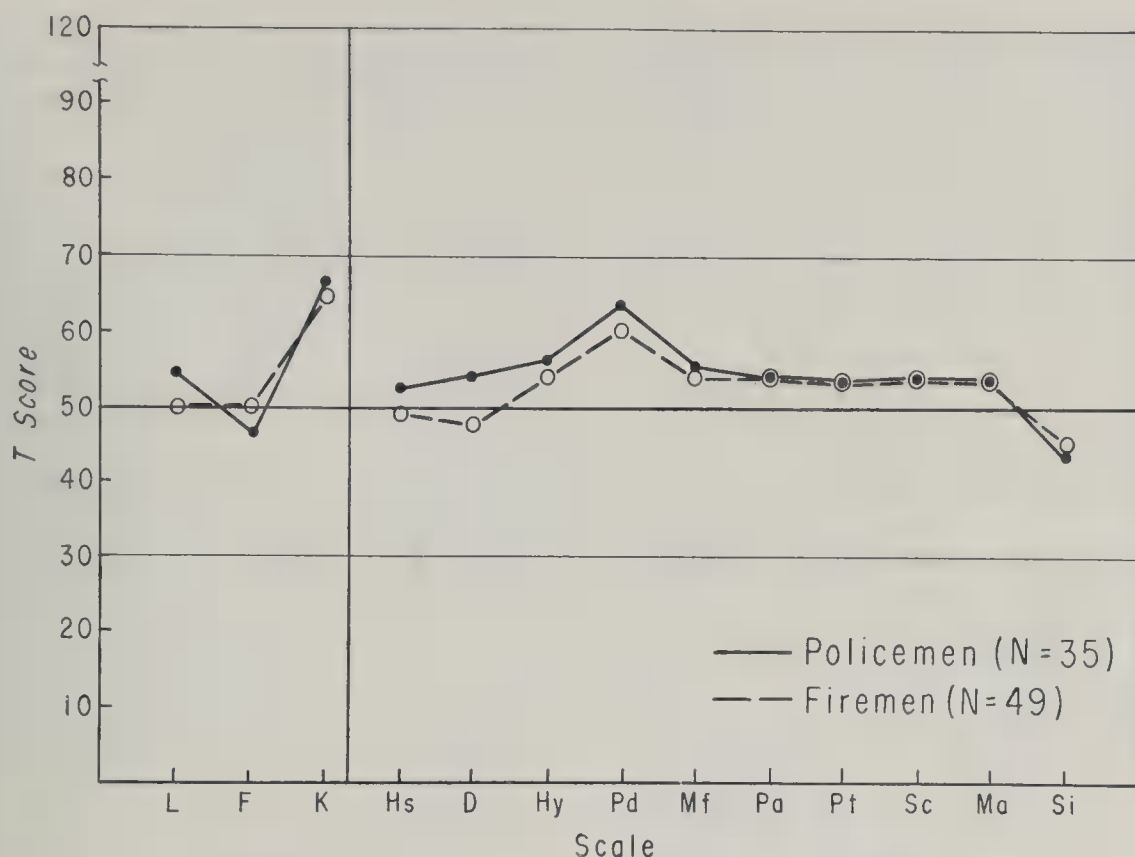


FIG. 4. MMPI profiles for policeman and fireman applicants.

Figure 4. Although the policemen and firemen applicant groups are small, the results in Figure 4 clearly show that: (a) the mean MMPI profile of policemen and firemen applicants is remarkably similar and (b) the mean profile of neither group contains any pathologically high scale scores (i.e., no mean above a *T* score of 70) indicative of serious psychopathology for the group as a whole. In general, the mean profiles shown in Figure 4 are not unlike those found in groups of male college students.

The elevated *K* scale, indicating defensiveness against admitting psychological weakness, is a function either of the stringent medical selection through which each applicant already had been processed or, more likely, an understandable cautiousness in a job-selection process. In terms of profile analysis, the elevations on *Pd*, *Hy* (43-code), and low *Si* (low 10) profile shown by both groups of applicants indicates that they are typical of the enlisted men one often encounters in the military services: blustery, sociable, exhibitionistic, active, manipulating

others to gain their own ends, opportunistic, unable to delay gratification, impulsive, and showing some tendencies toward overindulgence in sex and drinking (Dahlstrom & Welsh, 1960). In a word, fitting the lower socioeconomic group's stereotype of the "man's man."

DISCUSSION

The results with this 3-year sample indicate that the men on the eligibility lists for both policemen and firemen recruits in this large city in the Northwest are of surprisingly high intelligence. That this finding does not reflect a recent change is shown by the fact that the 1962 study by Zaice (1962), of 104 Portland policemen, mean age of 41.1 years, who had been on the force an average of 15.8 years, revealed that these experienced police officers earned a mean AGCT score of 126, standard deviation of 10. This mean score places them at the ninetieth percentile on the AGCT. The 49 patrolmen in the total of 104 studied by Zaice earned a mean AGCT score of 123

(eighty-fifth percentile), while the 27 detectives earned a score of 125 (eighty-ninth percentile) and 28 command personnel earned a score of 130 (ninety-third percentile). Interestingly, these results with the AGCT are surprisingly like those earlier reported by August Vollmer (unpublished mimeograph report) in a 1947 study of AGCT scores of policemen in the city of Portland.

Thus, whether measured by the WAIS, as by us, or the AGCT, as by Zaice (1962) and Vollmer (1947 unpublished), both experienced police officers and new recruits in this large city are of superior intelligence. While the Army Alpha Test findings on policemen of the early 1920s studied in a number of cities varied considerably from one city to the next (Dudycka, 1955, p. 59), those policemen being appointed in the past decade in cities other than Portland clearly are of superior intelligence; for example, the recent St. Louis patrolman recruit earns a mean AGCT score of 118, indicating superior (eightieth percentile) intelligence (DuBois & Watson, 1950), while policemen from all over the country examined during World War II also scored above average (Stewart, 1947).

These findings suggest that the 116 Portland policemen applicants who made the eligibility list and who were examined by us in the present study, although of superior intelligence, are not very different from the 1947 police officer on this same police force (Vollmer, 1947 unpublished), or the man currently on the force with some 15 years longevity (Zaice, 1962).

The recent study by Zaice permits two other comparisons of our newly eligible police recruit with the seasoned police officer now on the force. Figure 2 of the present study contains the mean EPPS personality profile of 93 of the 116 policemen applicants studied by us. Zaice (1962) also used the EPPS on 95 of his 104 experienced Portland police officers. A comparison of his results with ours shows an EPPS profile for these men on the force which is almost identical to that shown in Figure 2. Thus, our eligibility list recruit is like the more seasoned police officer on this

15-personality-need inventory also. Since Zaice also used the SVIB, this comparison also is possible. Examination of the results obtained by Zaice with 95 men and those obtained on our 112 police applicants shown in Figure 3 again revealed that the profiles were almost identical.

Zaice's study clearly can be considered to have provided us with validity correlates in that his results, obtained on police officers who have been in the Portland Police Department for over 15 years, permit a test of the degree to which current successful recruits resemble men already in police work in this city. The comparisons make clear that the 1959-62 police recruit is a remarkable facsimile (on at least three variables) of the police officer appointed to the same force in 1946-47. These results permit the guess that he probably is like his older counterpart on the other personality measures examined by us but not by Zaice.

Because no comparable study of men now in the fire Department has been done, one can only speculate that our 127 recruits are not too different from Portland firemen who have been on the job 15 or more years, although Stewart (1947) found that, based on a national sample, fire fighters scored slightly lower on the AGCT than did police patrolmen.

The finding that our sample of 243 men had a mean WAIS IQ of 112 (eightieth percentile) raises some questions regarding their educational attainment. Fourteen percent of our sample dropped out of school before high-school graduation, 46% completed only high school, and only 3% graduated from college. Yet the results shown in Figure 1 clearly indicate that most, if not all, of the 243 men probably could have completed college. The clinical interview with these men established that, coming as they did from our lower socioeconomic family backgrounds, these men, clearly of high intelligence, have failed to go to college for two basic reasons: (a) poor motivation to pursue their education, with realization of the importance of such education only after completing their military service or starting their own families; and (b) lack of financial resources

in the family, or lack of family recognition of the importance of schooling, and the consequent failure to provide the appropriate overt and covert values and incentives in the home which would stimulate or reward the educational process in these young men.

Thus, as a group, they represent men who, endowed with good intellectual abilities which they have not fully developed or tested, often in their own words turn to police work and firefighting as career choices next best to the business and other professions for which they qualify but for which their educational limitations, lack of knowledge, or lack of opportunity prevent them from entering. The result is that firefighting and police work are recruiting men of superior intelligence into their ranks; an important first requirement if these two occupations are to succeed in what may well be the attainment of professional status for one, or another, or both of them (Gourley, 1961).

REFERENCES

- ANIKEEF, A. M., & BRYAN, J. L. Kuder interest pattern analysis of fire protection students and graduates. *J. soc. Psychol.*, 1958, **48**, 195-198.
- CHENOWETH, J. H. Situational tests: A new attempt at assessing police candidates. *J. crim. Law Criminol. police Sci.*, 1961, **52**, 232-238.
- DAHLSTROM, W. G., & WELSH, G. S. *An MMPI handbook: A guide to use in clinical practice and research*. Minneapolis: Univer. Minnesota Press, 1960.
- DUBOIS, P. H., & WATSON, R. I. The selection of patrolmen. *J. appl. Psychol.*, 1950, **34**, 90-95.
- DUDYCHA, G. T. *Psychology for law enforcement officers*. Springfield, Ill.: Charles C Thomas, 1955.
- EDWARDS, A. L. *Manual for the Edwards Personal Preference Schedule*. (Rev. ed.) New York: Psychological Corporation, 1959.
- FROST, T. M. Selection methods for police recruits. *J. crim. Law Criminol.*, 1955, **46**, 135-145.
- GOURLEY, G. D. State standards for local police recruitment and training. Paper read at American Association for the Advancement of Science, Los Angeles, December 1961.
- HAGENAH, T. Normative data, patterning, and use of the Strong Vocational Interest Blank. In W. L. Layton (Ed.), *The Strong Vocational Interest Blank: Research and uses*. Minneapolis: Univer. Minnesota Press, 1960. Pp. 104-117.
- KATES, S. L. Rorschach responses, Strong Blank scales, and job satisfaction among policemen. *J. appl. Psychol.*, 1950, **34**, 249-254.
- KOLE, D. M. A study of intellectual and personality characteristics of medical students. Unpublished master's thesis, University of Oregon Medical School, 1962.
- MATARAZZO, J. D. The interview. In B. B. Wolman (Ed.), *Handbook of clinical psychology*. New York: McGraw-Hill, 1964, in press.
- MATARAZZO, R. G., MATARAZZO, J. D., & SASLOW, G. The relationship between medical and psychiatric symptoms. *J. abnorm. soc. Psychol.*, 1961, **62**, 55-61.
- MULLINEAUX, J. E. An evaluation of the predictors used to select patrolmen. *Publ. Personnel Rev.*, 1955, **16**, 84-86.
- OGLESBY, T. Use of emotional screening in the selection of police applicants. *Publ. Personnel Rev.*, 1957, **18**, 228-231.
- STERNE, D. M. Use of the Kuder Preference Record, Personal, with police officers. *J. appl. Psychol.*, 1960, **44**, 323-324.
- STEWART, NAOMI. AGCT scores of army personnel grouped by occupation. *Occupations*, 1947, **26**, 5-41.
- STRONG, E. K. *Vocational interests of men and women*. Stanford: Stanford Univer. Press, 1943.
- THORNDIKE, R. L., & HAGEN, ELIZABETH. *Ten thousand careers*. New York: Wiley, 1959.
- ZAICE, J. E. Measured interests, personality, and intelligence of professional policemen. Unpublished master's thesis, Washington State University, 1962.

(Received April 2, 1963)

PREDICTOR VARIABLES FOR CREATIVITY IN INDUSTRIAL SCIENCE

FRANCIS E. JONES

Human Factors Laboratory, Rensselaer Polytechnic Institute

25 managers rated the creative performance of a representative sample ($N = 88$) of industrial scientists and technologists within a large company. The ratings provided a continuous distribution of creativity scores. Performances on 53 different test variables were correlated with the criterion scores to produce 9 valid predictors of the rated criterion. The best multiple correlation obtainable upon the smallest number of the test predictors with rated creativity is .67, corrected for bias. A limited cross-validation upon a different sample confirmed the validation. According to test results, the typically more creative industrial scientist or engineer was found to be: highly capable of reasoning well with words and other symbols, fluent in the output of ideas, original in the quality of ideas, emotionally stable, determined to master his working environment, adventurous in outlook, high in degree of scientific curiosity, and low in indication of general anxiety.

Creativity is problem solution wherein a substantial component of uniqueness is involved. The creative problem may be almost any problem, and the uniqueness impartially may reside either in the problem formulation, or in the solution process, or in the resultant product. But however creativity is formally defined, there is the need in creativity research to identify relative degrees of creativity among individuals (in whatever field) by use of a "pointing" technique based on expert judgment; and to evaluate the measurable psychological characteristics which are found to be associated with the differing degree of creative behavior. If the program were executed for all fields, we would know, limited only by our capabilities of measurement, just which characteristics are generally predictive across fields, and also which ones are specific to certain fields only. The present study investigated creativity in chemists and chemical engineers and other scientists and technologists associated with them in an industrial setting.¹

METHOD

Problem

The specific problem of the research was to identify and to measure as a basis of prediction

certain paper and pencil patterns of response which are stably enough associated with larger creative systems of response to be of interest in the predictive sense. The approach followed consisted of two parts: A representative sample of industrial scientists and technologists was put into rank order for productive creativity, and the test performance of the sample within several areas of psychological measurement was analyzed to find stable predictors for creative success.

Sample Studied

The N of the validation sample was 88 cases. The subjects were male salaried employees in good standing at Naugatuck Chemical Division of the sponsoring company. These men were engaged in three different functions: research and development, 61%; production and miscellaneous, 29%; and sales and advertising, 10%. To the nearest year, their ages ranged from 26 to 56 with the median at 39, and their time with the company ranged from 1 to 30 with the median at 12. All subjects had at least the bachelor's degree or its equivalent, and a substantial proportion held advanced degrees (35%). By academic training, 44% were chemists, 36% were chemical engineers, and 20% were from other formal disciplines, such as: agriculture, business administration, literature, management engineering, mechanical engineering, natural science, psychology, and zoology. As measured on the California Short-Form Test of Mental Maturity, 28% were very superior in IQ (130-up), 49% were superior (115-129), 19% were high average (100-114), and 3 cases were low average (85-99).

was the source of initiation and major support; to more limited extent the research was supported by the Wayne Research Center.

¹ The research described in this report was supported by the United States Rubber Company. Within the company, Naugatuck Chemical Division

Criterion of Creative Performance

All subjects were rated on a 100-point scale on each of 12 descriptors. All ratings were made by managerial personnel. Before rating individual subjects, 25 managers rated each criterial descriptor for its relevance to productive creativity at Naugatuck. The procedure used in rating and weighting the descriptors was essentially that described by Sprecher (1959). The descriptors rated and the relative weights assigned to them by the 25 managers were as follows: Analytical Mindedness, 2; Communicativeness, 1; Idea Mindedness, 8; Level of Energy, 6; Liking for Problems, 8; Organization in Work, 1; Originality, 10; Perseverance, 2; Personal Relations, -2; Practical Mindedness, 3; Self Reliance, 3; Technical Competence, 6. Each rater was calibrated, on the basis of his ratings of the subjects, with a correction for each descriptor, so that without affecting his spread of ratings these ratings could be brought to scale levels comparable to those of all other raters. As ratings per subject ranged from one to nine in number, the mean of corrected ratings was determined, when appropriate for a given subject, before weight was applied to a descriptor. The sum of his weighted descriptors was the criterion score on creativity for each subject. The criterion scores thus obtained for the 88 cases produced a continuous distribution of creativity scores of satisfactory range. Converted to a 50-point scale, the range of scores was 11.9 to 43.7; the mean was 32.6; the sigma was 7.1. On the same scale, the median was 34.6; cases above the median were designated "more creative" and cases below the median "less creative."

Reliability of Criterion

In order to estimate the reliability of the criterion, several weeks after the composited ratings were completed for all subjects, the 25 raters were, unexpectedly to them, called upon further to provide a set of global ratings of creativity. These later ratings were made on a 100-point scale. They correlated .88 with the original ratings.

Test Variables Evaluated

Predictor variables were sought in measures of both aptitude and attitude. By combining forms of tests, and different but similar tests and subtests, scales of sufficient length were achieved for the measurement of 53 different variables with adequate expectancies of reliability. Each scale was intended to measure one definite factor, or at least a closely factorlike entity. The following tests composed the 2-day battery: California Short-Form Test of Mental Maturity; Culture Free Test of "g," Subtest IV *only* (Forms 3A and 3B); Ship Destination Test; Miller Analogies Test; Logical Reasoning Test; Doppelt Mathematical Reasoning Test; Minnesota Engineering Analogies Test; Rensselaer Test of Estimations; Word Fluency Test; Associational Fluency Test; Ideational Fluency

Test; Expressional Fluency Test; Pertinent Questions Test; Alternate Uses Test; Consequences Test; Rensselaer Test of Creative Thinking; Kuder Preference Record; 16 Personality Factors Questionnaire (Forms A, B, and C); and California Structured Objective Rorschach Test.

Experimental Design

All subjects were rated before any subjects were tested. For practical reasons, the 88 subjects could not all be tested at the same time. They were tested in three separate groups, over a total span of about 17 months from first to last testing. In each of the separate groups, the subgroups of "more" and "less" creatives were equated for *N*, and as nearly as possible for all other control variables. The mean differences in criterion scores between the subgroups "more creative" and "less creative" were not the same for the three separate test groups. As a deliberate feature of the design, the subgroup mean difference was largest in the case of the first group tested (17.2 on the 50-point criterion scale); smaller in the case of the second group tested (9.7); and smallest in the case of the third group tested (5.9). Thus the average difference in rated creativity between subgroups was made progressively to converge throughout the course of the validation. An accompanying effect of this design procedure was naturally a shrinkage of the test group variances (in criterion score) in the same order of progression, from first to last. The effect of the design was examined in the case of all test scales administered (53 scales). In order to qualify as a predictor, a test variable: had to differentiate throughout the three separate test groups with all mean differences (in test scores) between subgroups "more creative" and "less creative" being of the same sign (either plus or minus); and had to differentiate overall, when the respective subgroups were pooled, according to a point-biserial coefficient of correlation equal to .25 or greater. A limited cross-validation was

TABLE 1
PREDICTORS CORRELATED WITH CRITERION

Test variables	<i>r</i> with criterion
Aptitudinal:	
Logical Ratiocination (LR)	312
Mathematical Ratiocination (MR)	286
Ideational Fluency (IF)	325
Originality (ORIG)	281
Attitudinal:	
Emotional Stability (C)	254
Dominance (E)	364
Venturesomeness (H)	241
Experimenting Attitude (Q1)	353
Guiltproneness (O)	-307

TABLE 2
PREDICTORS CORRELATED WITH EACH OTHER

	LR	MR	IF	ORIG	C	E	H	Q1	O
LR	—	616	308	204	376	161	101	277	—256
MR	616	—	354	110	106	134	106	259	—242
IF	308	354	—	534	157	361	330	192	—064
ORIG	204	110	534	—	204	336	325	185	—072
C	376	106	157	204	—	096	292	063	—673
E	161	134	361	336	096	—	555	417	—328
H	101	106	330	325	292	555	—	243	—425
Q1	277	259	192	185	063	417	243	—	—390
O	—256	—242	—064	—072	—673	—328	—425	—390	—

Note.—See Table 1 for definitions of abbreviations.

based on a small stratified random sample drawn from another location of the company (the Wayne Research Center).

RESULTS

Out of the 53 test variables evaluated in this study, 9 proved to be valid according to the adopted criteria, and hence were selected as predictors of creativity. Four of the 9 were aptitudinal in nature, and 5 of the 9 were attitudinal (see Table 1).

All of the attitudinal predictors come from the 16 Personality Factor Questionnaire (Forms A, B, and C combined). Among the aptitudinal predictors: LR consists of California Short-Form Test of Mental Maturity, Subtests 3 and 4, combined with Logical Reasoning Test; MR consists of Doppelt Mathematical Reasoning Test combined with Minnesota Engineering Analogies Test; IF is Ideational Fluency Test; and ORIG consists of Consequences Test combined with Rensselaer Test of Creative Thinking. The correlations of predictors with each other appear in Table 2.

All of the correlations shown, in both tables, are Pearson *r* coefficients. The best multiple coefficient obtainable upon the smallest number of predictors is .67, corrected for bias. The matrix solved for this coefficient

is made up of the *r*'s in Table 2 plus biserial coefficients with the criterion (not shown). The predictors used are LR, MR, ORIG, E, Q1, and O.

For a limited cross-validation of the above results, 10 subjects from the Wayne Research Center were tested with the battery used at Naugatuck. These 10 subjects were drawn randomly from pools of subjects previously rated either more creative (above the guessed median of the Center) and less creative (below that median). The selection was made so as to result in equal subgroups (*N*s of five and five). Test predictions of creativity were made on the 10 cases by use of Naugatuck multiple-regression weights applied to their raw predictor scores, with these products being summed in each case to produce a creativity score. The rank order of the creativity scores divides the cases at the median exactly in accord with the rating division. These predictions were forwarded before the Center ratings were ascertained.

REFERENCE

SPRECHER, T. B. A proposal for identifying the meaning of creativity. In C. W. Taylor (Ed.), *Third Research Conference on the Identification of Scientific Talent*. Salt Lake City: Univer. Utah Press, 1959. Pp. 29-45.

(Received April 8, 1963)

KENNETH E. CLARK, Editor
University of Rochester

Table of Contents

Color versus Shape Coding in Information Displays.....	
.....Sidney L. Smith and Donald W. Thomas	137
Effects of Variations in Rating Scale Format on Judgment.....	
.....Joseph M. Madden and Roger D. Bourdon	147
Sensory Feedback Analysis of Stereotelevision Pursuit Tracking.....	
.....John D. Gould and Karl U. Smith	152
Some Determinants of Job Satisfaction: A Study of the Generality of Herzberg's Theory...	
.....Robert B. Ewen	161
Predicting Success in Business.....	
.....Frank J. Williams and Thomas W. Harrell	164
A Factor Analysis of Retail Credit Application Data.....	
.....James H. Myers	168
Quantification of Biographical Data for Predicting Vocational Rehabilitation Success.....	
.....Raymond A. Ehrle	171
Effects of Different Patterns on Outcomes of Problem-Solving Discussion.....	
.....John K. Brilhart and Lurene M. Jochem	175
Voluntary Female Clerical Turnover: The Concurrent and Predictive Validity of a Weighted Application Blank.....	
.....William D. Buel	180
Remote Associates Test as a Predictor of Creativity in Engineers.....	
.....Lois-Ellin Datta	183
A Note on the Remote Associates Test, United States Culture, and Creativity.....	
.....Lois-Ellin Datta	184
The Comprehensibility of Several Grammatical Transformations.....	
.....E. B. Coleman	186
A Cross-Cultural Study of Achievement Motivation.....	
.....Harrison G. Gough	191
A "Contingent-Item" Method for Constructing a Short Personality Questionnaire.....	
.....Frank B. McMahon, Jr. and Raymond G. Hunt	197

American Psychological Association

Consulting Editors

GEORGE E. BRIGGS, *Ohio State University*
NORMAN FREDERIKSEN, *Educational Testing Service*
LEONARD D. GOODSTEIN, *University of Iowa*
EDWIN R. HENRY, *Peace Corps*
JOHN HOLLAND, *American College Testing Program*
CLIFFORD E. JURGENSEN, *Minneapolis Gas Company*

LAURENCE S. MCGAUGHRAN, *University of Houston*
QUINN MCNEMAR, *Stanford University*
HAROLD F. ROTHE, *Beloit Corporation*
THOMAS A. RYAN, *Cornell University*
ALEXANDER G. WESMAN, *Psychological Corporation*
CLARK L. WILSON, *Batten, Barton, Durstine, and Osborn, Incorporated*

This journal gives primary consideration to original investigations in any field of applied psychology except clinical psychology, although a descriptive or theoretical article may be accepted if it represents a special contribution in an applied field. Quantitative investigations of interest or value to psychologists working in the following broad fields will be considered: vocational and educational prognosis, diagnosis, and guidance at the secondary and college level; personnel research in business, industry, and government; engineering psychology; industrial working conditions; research on opinion and morale factors; job analysis and classification research; market and advertising research.

Manuscripts must be accompanied by an abstract of 100–120 words typed on a separate sheet of paper. The abstract should conform to the style of *Psychological Abstracts*. Detailed instructions for preparation of the abstracts appeared in the *American Psychologist* (1961, 16, 833), or they may be obtained from the Editor or from the APA Central Office.

Manuscripts should be addressed to the Editor:

Dr. Kenneth E. Clark
College of Arts and Science
302 Morey Hall
University of Rochester
Rochester, N. Y. 14627

All manuscripts must be submitted in duplicate. Original figures are prepared for publication; duplicate figures may be photographs or pencil-drawn copies.

Manuscripts must conform to the style requirements described in the *Publication Manual of the American Psychological Association*.

The following policies govern reprinting of materials copyrighted in APA journals: (a) to require approval to reprint for tables and figures, and for text only if more than 500 words [total from one article] in length; (b) to continue approval to reprint, contingent on the author's approval, for articles reprinted in whole or in major part; (c) to negotiate where possible the dedication of royalties from commercial publishers to the American Psychological Foundation.

Journal of Applied Psychology

Published bimonthly by the
American Psychological Association
Prince and Lemon Sts., Lancaster, Pa. 17604
and 1333 Sixteenth St. N. W.
Washington, D. C. 20036

\$10.00 per volume

\$2.00 per issue

HELEN ORR
Managing Editor

ELIZABETH S. REED
Advertising Manager

FRANCES L. BREWER
Editorial Assistant

Subscriptions, orders, and business communications should be addressed to the American Psychological Association, 1333 Sixteenth St. N. W., Washington, D. C. 20036. Address changes must reach the subscription office by the tenth of the month to take effect the following month. Undelivered copies resulting from address changes will not be replaced; subscribers should notify the post office that they will guarantee second-class forwarding postage. Other claims for undelivered copies must be made within four months of publication.

Second-class postage paid at Lancaster, Pa.

© 1964 by the American Psychological Association, Inc.

Journal of Applied Psychology

VOL. 48, No. 3

JUNE 1964

COLOR VERSUS SHAPE CODING IN INFORMATION DISPLAYS¹

SIDNEY L. SMITH AND DONALD W. THOMAS

MITRE Corporation, Bedford, Massachusetts

8 Ss counted objects of a specified color or shape on displays of 20, 60, or 100 items. Counting time and errors increased with increasing display density. Counting based on a 5-valued color code was faster and more accurate than counting using any of 3 shape codes. Color counting was not affected by the particular shape code on which the colors were superimposed. Shape counting was somewhat faster and/or more accurate when color did not vary on the display, and vice versa. Differences in counting performance appeared among the 3 shape codes and among certain of the symbols within shape codes, and small differences were confirmed among the particular code colors used.

Published research pertinent to display color coding has been summarized recently in a review article by Jones (1962). Of particular interest are several experimental studies which indicate that color coding is a powerful means of enhancing discriminability among classes of items presented in unstructured visual displays. Two such studies, one by Green and Anderson (1956) and a later replication and expansion of this work by Smith (1962), illustrate the potential value of redundant display color coding for visual search tasks. A third study (Smith, 1963a) confirmed that the value of redundant color coding is consistent in both visual search and class counting tasks. A logical extension of this work is to examine the use-

fulness of color as a nonredundant code, and to compare it in this regard with other possible visual coding dimensions, in particular with shape coding.

Some previous studies cite results which suggest that color coding might be superior to shape coding under certain conditions. Eriksen (1952), in the context of a search task, concluded that visual separability based on seven hues was better than that provided by a symbol set of seven geometric forms. More recently, Hitt (1961) discovered a code of eight colors to be superior to eight-valued letter, geometric shape, and configuration codes, but equivalent to a numerical code, in locating, counting, comparing, and verifying tasks. He found color to be inferior in an identifying task. Christner and Ray (1961), using slightly different eight-valued codes in a different experimental context, concluded that color was superior to numeral and shape codes in locating and counting tasks, but inferior to numerals in an identifying task.

Coding based on shape or symbology is, of course, a most versatile means of presenting information in visual displays—because of the large “alphabet” of discriminable symbols that can be used, because prior experience of the viewers can often be relied upon for interpretation of symbols, and because

¹ The authors wish to acknowledge the help of David E. Moore, who assisted in data collection for the experimental study described in this report, and of Barbara B. Farquhar, who assisted in the data analysis. The research reported in this paper was sponsored by the Air Force Electronic Systems Division, Air Force Systems Command, under Contract AF19(628)2390. This paper will also be obtainable as ESD Technical Documentary Report Number 63-483. Further reproduction is authorized to satisfy needs of the United States Government. The MITRE Corporation has obtained approval for release of the information contained in this report. A more detailed account of this research is available from the authors (Smith & Thomas, 1963).

sequences of symbols can be used as a language to increase still further their effectiveness in conveying information. For these reasons, a display designer would not generally be able to replace a symbol code with a color code even if he wished to do so. However, when the number of critical classes of displayed items is relatively small, as in the studies cited above, the selection of a color rather than a symbol code might well be preferable. This present study represents an attempt to measure systematically this superiority of display color coding, by comparing it with various shape codes in the context of a relatively simple operator task, that of counting a particular class of displayed items.

PROCEDURE

Eight men and women with normal color vision participated as subjects (Ss) in this study. Each S was run individually, completing 550 trials distributed over four experimental sessions. In each trial Ss were required to count the number of items of a predesignated "target" class that were displayed under varying conditions. Counting times and errors, if any, were recorded.

Displays were prepared as 2×2 inch color slides and presented by rear projection on a large viewing screen to produce a 29-inch square display field. The Ss were seated approximately 5 feet from the screen, with the center of the display field at eye level. Ambient illumination was moderately high, about 2 foot-candles.

Each of the projected displays consisted of 20, 60, or 100 symbols, bright-colored figures on a dark background. The symbols were presented in random² arrangement on the displays, in a set of positions chosen from 400 possibilities that can be imagined as representing the intersections of a 20-column by 20-row partitioning. Different random positions were used for different displays, and the particular symbol and color displayed at each position were also chosen randomly, with certain restrictions that will be noted.

Three sets of code symbols were used in this study. Each set contained five symbols. Figure 1 illustrates the various display symbols used.³ Some displays

were prepared using aircraft shapes (silhouettes). These were chosen to represent a fairly difficult shape code, in that each shape was similar to at least one other. Other displays were prepared using a set of conventional geometric forms. A final group of displays were based on a set of military symbols that were chosen to represent an easily discriminable code: the symbols differed from one another in size and general orientation as well as in specific shape.

Some of the displays used were multicolored. Each symbol could appear in any of five colors—green, blue, white, red, or yellow—randomly chosen. In other displays, the symbols were shown in just one of the five colors. Three observers agreed that the colors as projected could be described adequately by the Munsell color matches listed in Figure 1.

As noted already, the number of times a particular symbol, or color, appeared on any one display was predetermined on a random basis. The expected frequency would be one fifth the total display-item density, but the actual numbers could vary from this figure. That is to say, in a display of 100 geometric forms, there might not be 20 triangles, but instead 23, or even 30. The display slides, however, were prepared in groups of five, and in each such group the overall frequency of each symbol or color *was* equated: In the example just cited, there would be a total of 100 triangles in the five-slide set of 100 density geometric-form displays, as well as 100 circles, 100 red symbols, etc.

Altogether, 65 display slides were used, comprising three subsets. A basic set of 45 slides was made up of five displays at each of the three density levels for each of the three shape codes, in which colors were distributed randomly with respect to shape. That is to say, in a display of any particular symbol class, one could expect to find some symbols of each of the five types in any of the five colors. Ss counted each of these displays 10 times in the course of the study, once for every shape and once for every color.

A second set of slides was used to examine the effectiveness of shape coding when color did *not* vary concomitantly. It was made up of fifteen 100-item displays that were single colored, one of each color for each of the three symbol classes. The Ss counted each of these displays five times during the study, once for every shape.

A third set of slides was used to ascertain the effectiveness of color coding when shape was held constant. This set consisted of five multicolored displays, on each of which one particular military symbol appeared 100 times, represented in all of the various colors, with one display slide for each of the five symbols. The Ss counted each display of this set five times, once for every color.

The order of presentation of various displays, as well as the choice of target class from trial to trial, was randomized separately for each S. Before each

² The word "random," as used in this report, is intended in every instance to indicate an unbiased selection from among equiprobable alternatives, with any qualifying restrictions indicated as such.

³ The display symbols used were chosen from those available commercially from drafting supply firms, specifically Chart-Pak, Inc. (Leeds, Mass.); Mico/Type, Inc. (Los Angeles, Calif.); and Para Tone, Inc. (La Grange, Ill.). Each display was made initially by placing black symbols on a white background. From this, a large photonegative was prepared, and trans-

parent colored tape was affixed over the various display items as appropriate, after which this display was reduced photographically to slide size using transmitted light.
















COLORS (MUNSELL NOTATION)	MILITARY SYMBOLS	GEOMETRIC FORMS	AIRCRAFT SHAPES
GREEN (2.5 G 5/8)	RADAR 	TRIANGLE 	C-54 
BLUE (5 BG 4/5)	GUN 	DIAMOND 	C-47 
WHITE (5 Y 8/4)	AIRCRAFT 	SEMICIRCLE 	F-100 
RED (5 R 4/9)	MISSILE 	CIRCLE 	F-102 
YELLOW (10 YR 6/10)	SHIP 	STAR 	B-52 

FIG. 1. Color and shape codes used. (For consistency, military designations are used to denote the aircraft shapes. As actually projected, the diameter of the circle was $\frac{1}{2}$ " on the screen, with the other symbols in proportion as shown.)

trial, a sample symbol or color patch was presented on a small-auxiliary display device that was placed on a table between the *S* and the large display screen, indicating to *S* the particular target class that was to be counted. The large display was then exposed to begin the trial which ended when the *S* announced his count. The experimenter recorded this count, and the time required, and initiated the next trial, choosing a new display and target symbol. Each *S* was instructed to work as quickly and accurately as possible, leaving it to his individual judgment as to what constituted acceptable performance. The *Ss* were given no feedback as to the accuracy of their counts during the course of the experiment.

RESULTS

Average counting time using the different codes is plotted in Figure 2 as a function of the number of displayed items. Counting based on the color code resulted in average

times that were identical for all three symbol classes on which the colors were superimposed, so these data have been combined and are shown as a single-average curve. It can be seen that as the number of displayed items becomes larger, counting time increases linearly at a characteristic rate for each display code used. These four functions, if extrapolated, would intercept the ordinate (for zero display density) at about 1 second, presumably reflecting the verbal reaction time for this particular group of *Ss*. From inspection of these data it is evident that colors were counted about twice as fast as the best set of symbols and three times as fast as the poorest symbol code.

Figure 3 illustrates a corresponding increase

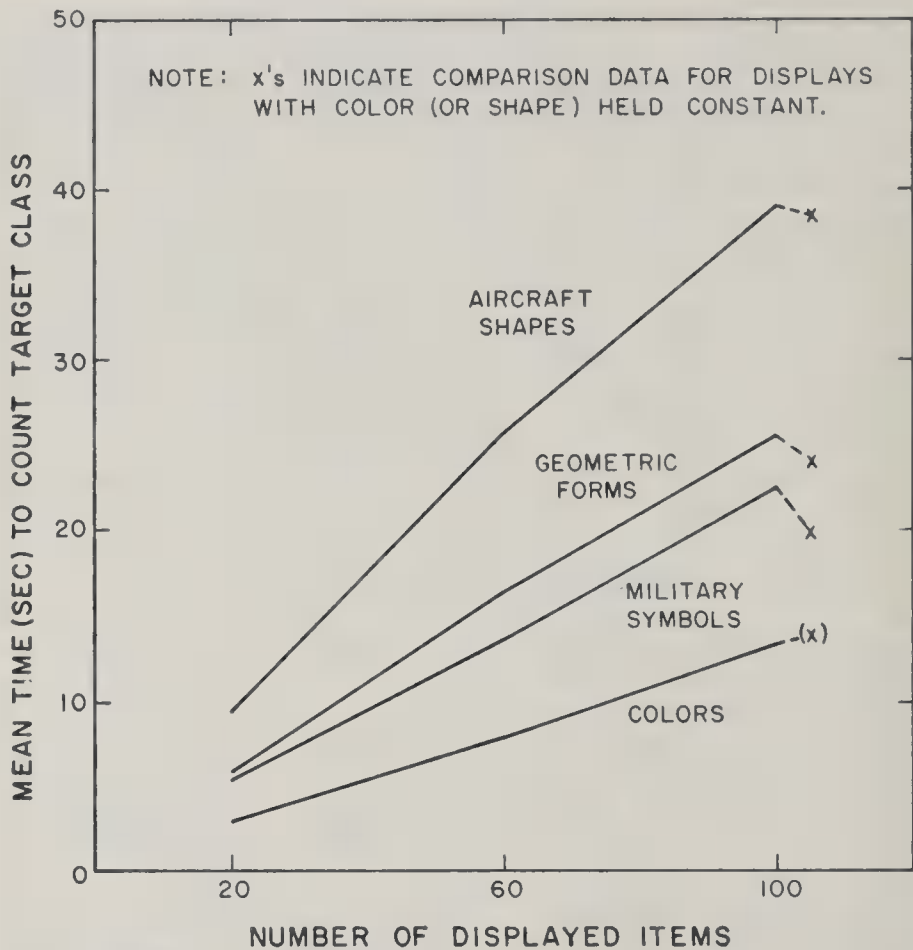


FIG. 2. Average counting time as a function of display density, comparing color coding with the three shape codes.

in frequency of counting errors with increasing display density. There was no significant difference in error frequency for color counting related to the particular shape code on which the colors were superimposed, and these data are combined in Figure 3 to show just one curve. These results tend to confirm the conclusions based on the time data already reported, i.e., fewer errors were made in color counting than in shape counting, and fewer errors were made at low display densities than at high. Analysis of counting errors in terms of their sign (most were underestimates) or size indicated the same trends as the results based simply on error frequency shown here.

The design of this study permits a number of statistical comparisons of the two performance measures, counting time and errors, under different experimental conditions. Statistical treatment of counting time was based on analysis of variance for time scores

summed in different ways across trials depending upon the particular comparison involved. The approach used followed that advocated by Edwards (1950) for the case of repeated measurements from the same Ss. Statistical treatment of counting errors was based simply on chi-square analysis of the relative frequency of error trials and correct counts.

The first comparison of interest in the data analysis is an overall comparison between color and shape counting, at different display-density levels and for different shape codes, where color and shape both vary on the displays. Analysis of variance, for this comparison, was based on the sum, for each S, of counting times for 25 trials, 5 trials for each of the 5 colors or symbols in the code. Statistically reliable differences in counting time were confirmed ($p < .001$) attributable to display density, the particular shape code displayed, the code used for counting (whether

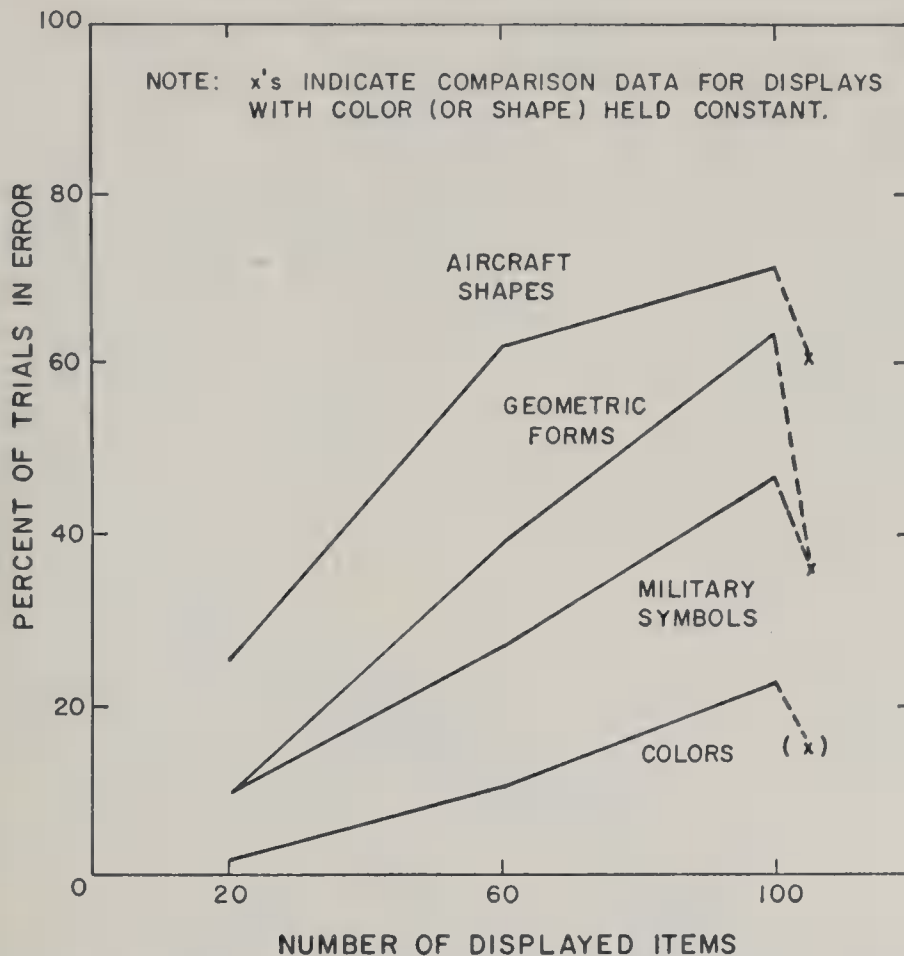


FIG. 3. Counting errors as a function of display density, comparing color coding with the three shape codes.

color or shape), and all interactions of these variables. Supplementary range tests confirmed that each level of increasing display density resulted in a significant increase in counting time, that the aircraft shape code was significantly different from the military symbol and geometric-form codes ($p < .001$), and that these latter two codes were probably different from one another ($p < .01$). Chi-square analysis of error frequency, for the same comparison of experimental conditions, also confirmed differences among display densities, among shape codes, and between color and shape counting.

A second comparison of interest involves counting times for target classes of different colors, for displays at different density levels, using different shape codes, where shape and color both vary. (This comparison is similar to that already described, except that the shape counting data is ignored and the partic-

ular target color is taken into account.) The scores used, in this and subsequent variance analyses, were the sums of counting times for five trials. The results confirmed significant differences ($p < .001$) attributable to display density, to target color, and to the interaction of these variables. There was *no* noticeable effect on color counting attributable to the shape code on which it was superimposed. A supplementary range test confirmed that green targets were the most difficult, followed by blue, with white, red, and yellow targets equivalent to one another in terms of counting time. Chi-square analysis of error frequency confirmed differences among display densities, and among target colors.

Figure 4 illustrates the differences among colors in average counting time. These differences are relatively small, which might tend to discount their potential importance for practical display applications. Moreover, al-

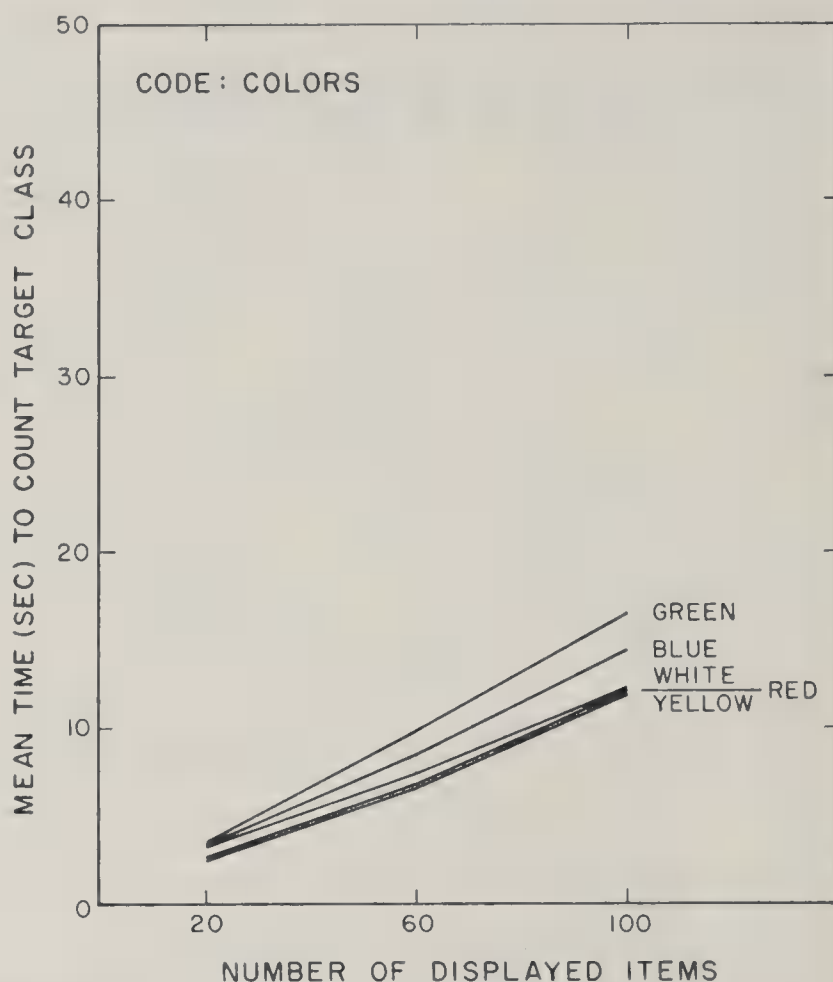


FIG. 4. Average counting time as a function of display density, comparing different color target classes.

though error differences among colors were confirmed, as noted above, it should be stated that in this case error frequency did not correspond exactly with the ordering of colors based on time scores.

A third comparison of interest is one involving counting particular colors, on 100-density military-symbol displays where the shapes varied, and on the comparable displays where shape was held constant. Variance analysis of time scores again confirmed differences among target colors ($p < .001$), with an ordering of colors similar to that in the previous analysis. There was no significant difference in counting time related to whether just one or several shapes were present in a display. However, the error-frequency analysis *did* indicate a statistically reliable difference ($p < .01$) attributable to variable shape. This refers to the difference between the last two points on the color curve in Figure 3.

Just as it was possible to confirm differences in counting performance among certain of the particular colors in the color code, differences can also be noted among some of the symbols in the shape codes. The time curves for counting each of the displayed symbol classes, as a function of display density, are presented for the military symbol code in Figure 5, for the geometric forms in Figure 6, and for the aircraft shapes in Figure 7. The relevant data analyses confirmed reliable differences ($p < .001$) in both time and error among the symbols within the military and geometric codes, and indicated there may also be differences in time ($p < .01$) among the aircraft shapes.

A final comparison of interest again involves shape counting, this time the counting of particular symbols in displays of 100 density when color was either varied or constant. The relevant statistical analyses again con-

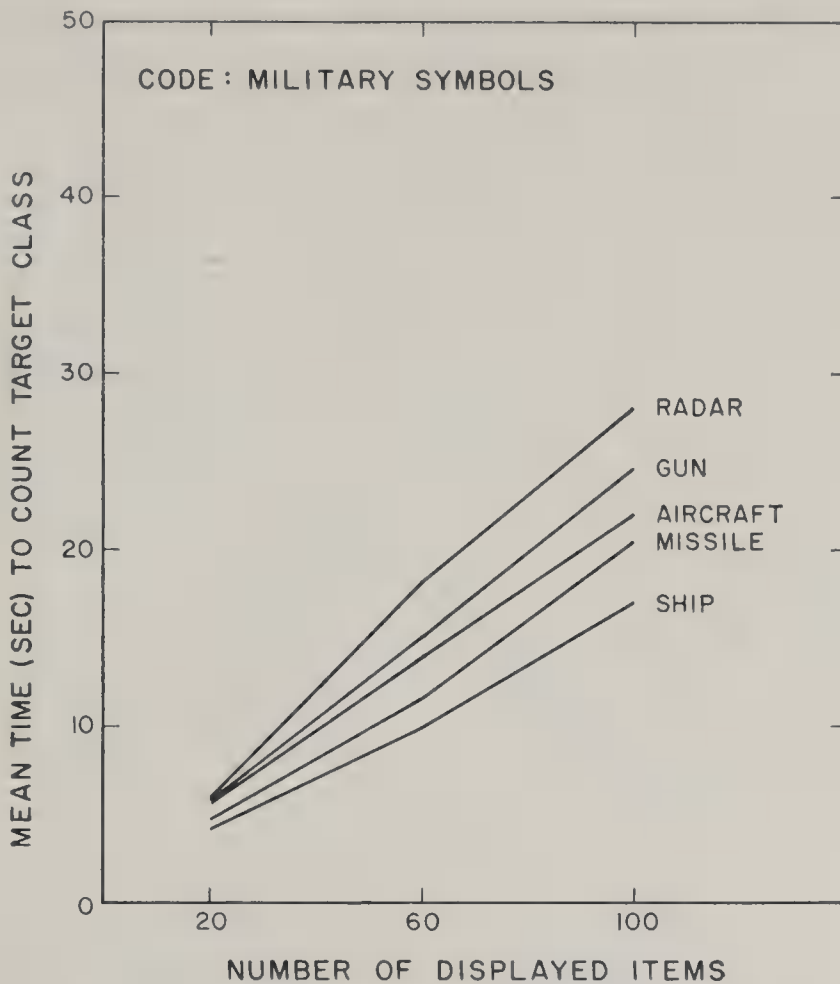


FIG. 5. Average counting time as a function of display density, comparing different military symbol target classes.

firmed significant differences ($p < .001$) in both time and error among symbols within the military and geometric codes. The analyses also confirmed ($p < .001$) or suggested ($p < .05$) differences in time or error or both between multicolored and single-colored displays, for each of the shape codes. This relates to the apparent improvement in speed and accuracy of shape counting when variable color was eliminated from the displays, indicated by the final data points plotted for the various shape codes in Figures 2 and 3.

Examining briefly the question of individual differences, rank-order correlation between overall average color-counting time and shape-counting time, paired for each S , indicates a very high correspondence ($\rho_s = .98$, $p < .001$). This suggests that each S benefited proportionally when counting colors rather than shapes. Various other possible rank-order cor-

relations based on overall individual performance turned out *not* to be statistically reliable: between shape-counting time and errors, $\rho_s = -.12$; between color-counting time and errors, $\rho_s = .35$; between color-counting errors and shape-counting errors, $\rho_s = .39$.

DISCUSSION

The linear relation between counting time and display density observed in this study is similar to that reported in an earlier study by Smith (1963a), who argued that the effect of a display code can be described economically in terms of the characteristic slope of this linear function. This resemblance of Smith's earlier results to those obtained in the present study is very striking: Despite differences in experimental procedure and the use of different S s, the average color-counting time curves, and error-frequency curves, are iden-

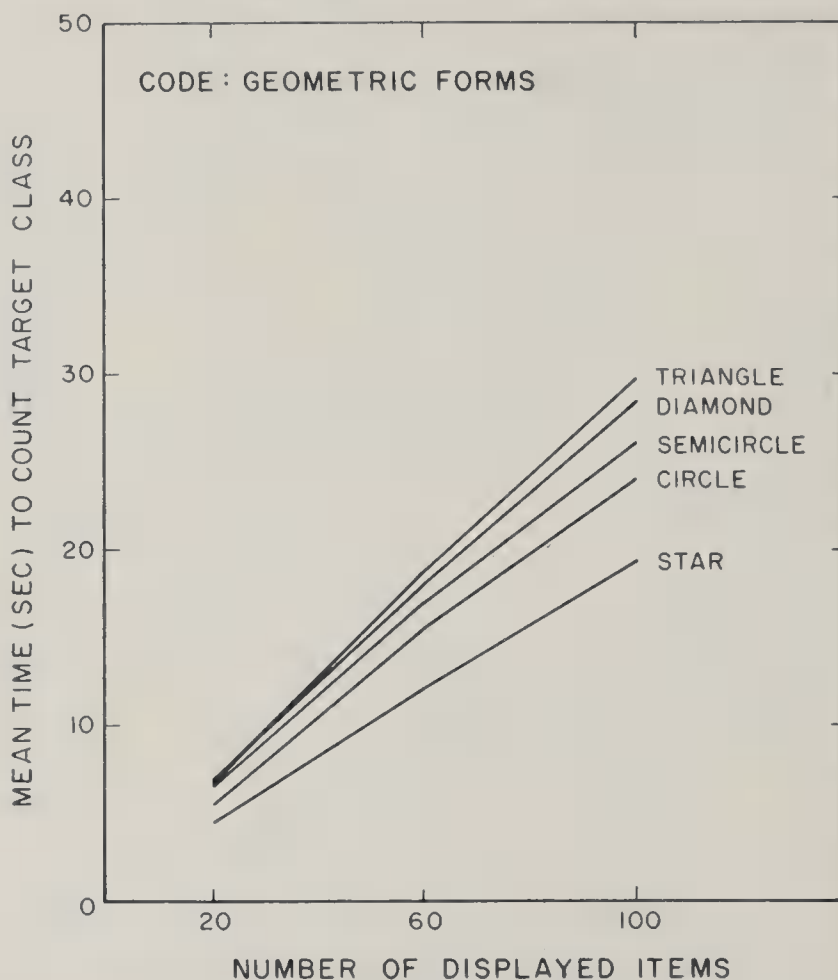


FIG. 6. Average counting time as a function of display density, comparing different geometric form target classes.

tical in the two experiments. This similarity of results tends to confirm the internal data in the present study which indicate that the effectiveness of color coding was independent of the particular shape code on which it was superimposed.

The overall results of the present study explicitly confirm the effectiveness of color coding for a counting task, in line with findings from previous research already cited. Moreover, the present data provide quantitative measures of the relative superiority of color over various shape codes. From the practical viewpoint of a display designer, this superiority should be taken into account in display organizing. The present results indicate that when color and shape are superimposed on a display, the color code will prove dominant in the visual separability it provides. Under these circumstances, color should be relied upon as

the primary means of coding displayed data relative to a user's task-information requirements, and shape coding should be used to denote secondary distinctions among displayed data.

It should not be argued from these results, however, that this superiority of color to shape would necessarily be maintained if one were to choose a larger display alphabet, a greater number of code categories, in each case. As more colors were added to a display, there would probably be a decrease in average discriminability that would eventually reduce the effectiveness of the color code, whereas adding more symbols to a shape code might not have an equal effect. Published studies, such as those by Halsey and Chapanis (1951), and Conover (1959), indicate that the limit of absolute discrimination for colors falls somewhere in the range of 5 to 12 alternatives

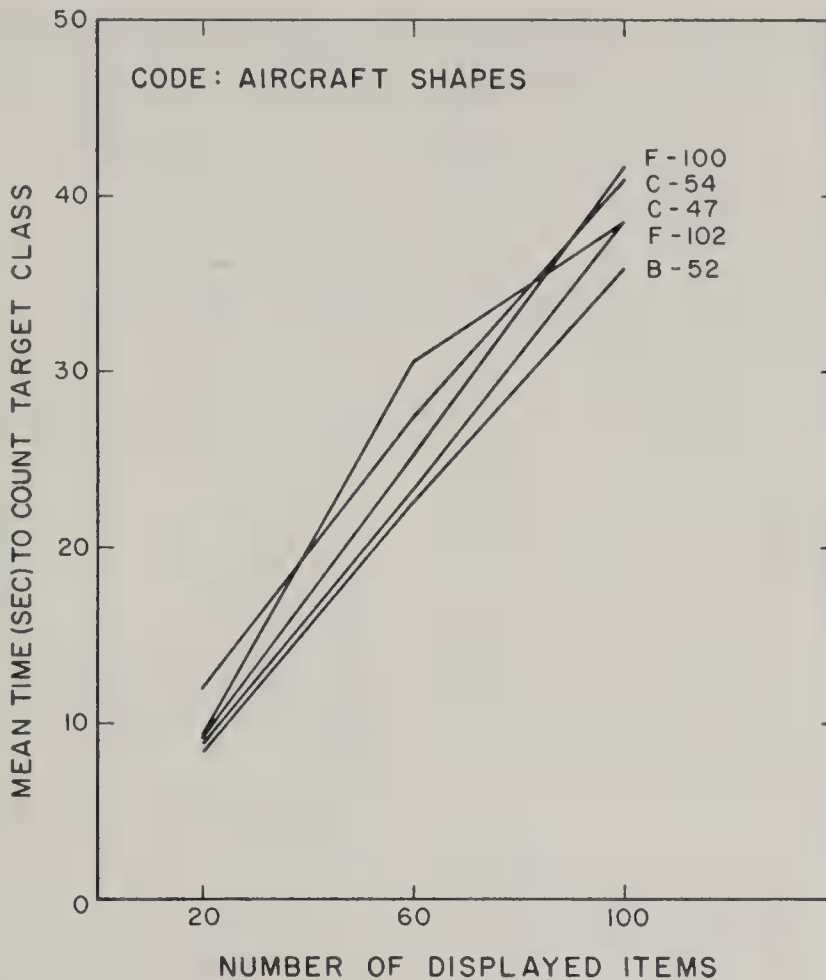


FIG. 7. Average counting time as a function of display density, comparing different aircraft shape target classes.

depending upon viewing conditions. Although it is probable that Ss could do somewhat better than this given special training under laboratory conditions, it is certain that in practical display situations a color code cannot provide a discriminable alphabet of a size that even approximates that available with symbol coding. It is for the particular case of a small number of display categories, as in this present study, that the clear advantage of color coding can be confirmed in the laboratory and exploited in display design.

In addition to this direct comparison of color with shape coding, there is another question of interest, namely, what are the effects on display utilization of *superimposing* color and shape coding. Here the question is not whether one should choose either color or shape coding, but rather what are the effects of using both together. One of the advantages

of color coding that is sometimes cited to advocate its use is that a color code can be added nonredundantly to a pre-existing shape code to carry additional information. This argument is made explicit by Anderson and Fitts (1958) who report results of research on human-information transmission to support the argument that "the most feasible means for increasing the size of a coding alphabet is to increase the dimensionality of the stimulus." As an alternative display-design approach, one might choose to add colors while reducing the size of the symbol code, without increasing the total amount of displayed information. This latter method was used in a study by Newman and Davis (1962) whose reported results seem to support the conclusion that the use of two or three colors in this nonredundant fashion improved both speed and accuracy in searching and decoding tasks.

Further clarification of the effects of color-shape superposition is needed and is, to some extent, provided in this present study, where the results indicate a measurable performance decrement in counting based on one code when it is superimposed on the other. This penalty may be more than outweighed by the advantages of such color-shape superposition in many practical display applications, but it should be taken into account. In this connection, it may be noted that Smith (1962) found no such detrimental effect of nonrelevant color coding on visual-search time for 3-digit numbers. Although there are a number of differences between this earlier work and the present study, there may be an indication here that counting provides a somewhat more sensitive performance index than searching. It is possible that this greater sensitivity results from the fact that a counting task provides the *S* with no explicit criterion of successful completion.

It is interesting that the differences confirmed among code colors, such as they are, follow approximately the ordering of color dominance reported in another study by Smith (1963b) dealing with the special question of legibility of colored numbers under conditions of overprinting. Under those conditions, white, orange and red symbols were more legible than green and blue. On the other hand, no such color differences were found for visual search tasks (Smith, 1962, 1963a), or for a counting task (Smith, 1963a) under conditions which yielded results otherwise identical to those of the present study.

The intracode differences noted among symbols may be regarded as fortuitous, since one need not necessarily expect differences in discriminability within a shape code. It should be emphasized also that the discriminability of a particular shape that permits its rapid counting is not necessarily a characteristic of that symbol, per se, but rather a reflection of how distinctive it is relative to the other symbols with which it is displayed. As an example of this, the aircraft symbol used in the mili-

tary code was counted more quickly and accurately than any of the particular symbols in the aircraft shape code. This is a phenomenon familiar to practical display designers, who have discovered it is often a wise precaution to verify empirically the discriminability of a particular symbol set proposed for use rather than to rely on data gathered in some different display context.

REFERENCES

- ANDERSON, NANCY S., & FITTS, P. M. Amount of information gained during brief exposures of numerals and colors. *J. exp. Psychol.*, 1958, **56**, 362-369.
- CHRISTNER, CHARLOTTE A., & RAY, H. W. An evaluation of the effect of selected combinations of target and background coding on map-reading performance. *Hum. Factors*, 1961, **3**, 131-146.
- CONOVER, D. W. The amount of information in the absolute judgment of Munsell hues. *USAF WADC tech. Note*, 1959, No. 58-262.
- EDWARDS, A. L. *Experimental design in psychological research*. New York: Rinehart, 1950.
- ERIKSEN, C. W. Location of objects in a visual display as a function of the number of dimensions on which the objects differ. *J. exp. Psychol.*, 1952, **44**, 56-60.
- GREEN, B. F., & ANDERSON, LOIS K. Color coding in a visual search task. *J. exp. Psychol.*, 1956, **51**, 19-24.
- HALSEY, RITA M., & CHAPANIS, A. On the number of absolutely identifiable hues. *J. Opt. Soc. Amer.*, 1951, **41**, 1057-1058.
- HITT, W. D. An evaluation of five different abstract coding methods. *Hum. Factors*, 1961, **3**, 120-130.
- JONES, MARI R. Color coding. *Hum. Factors*, 1962, **4**, 355-365.
- NEWMAN, K. M., & DAVIS, ANNE R. Non-redundant color, brightness, and flashing rate encoding of geometric symbols on a visual display. *J. engng Psychol.*, 1962, **1**, 47-67.
- SMITH, S. L. Color coding and visual search. *J. exp. Psychol.*, 1962, **64**, 434-440.
- SMITH, S. L. Color coding and visual separability in information displays. *J. appl. Psychol.*, 1963, **47**, 358-364. (a)
- SMITH, S. L. Legibility of overprinted symbols in multicolored displays. *J. engng Psychol.*, 1963, **2**, 82-96. (b)
- SMITH, S. L., & THOMAS, D. W. Display color coding compared with three shape codes for a class counting task. *MITRE tech. ser. Rep.*, 1963 (Dec.), No. 12.

(Early publication received December 4, 1963)

EFFECTS OF VARIATIONS IN RATING SCALE FORMAT ON JUDGMENT¹

JOSEPH M. MADDEN AND ROGER D. BOURDON

Personnel Research Laboratory, Lackland Air Force Base, Texas

The purpose of this study was to determine whether mean occupation evaluation ratings would differ as a function of 7 variations in rating-scale format. 60 basic airmen rated 15 occupations on 9 occupation-requirement factors for each format. A 3-way analysis of variance (occupations, factors, scale format) resulted in statistically significant terms for each of the main effects and for all 4 interaction terms. It was concluded that rating-scale format was a determiner of the judgment of raters in this sample and that selection of an optimal format should be based upon capability to predict a criterion.

Many personnel actions are based upon human judgment. For instance, personnel evaluation and occupation evaluation are illustrative of important areas of personnel management in which the only available measuring stick is judgment. In order to improve measurement in such cases where no physical scale is available, it is important that the methods and procedures used to obtain judgments result in a maximal degree of objectivity. Conditions which distort judgment and lead to invalid results must be identified and eliminated from the judgmental situation.

A series of studies have been undertaken to specify the dynamics of judgment in the occupation-evaluation situation as a part of the development of improved occupation-evaluation procedures for the Air Force. This work has dealt with the effects of the context in which an occupation is evaluated (Madden, 1960a, 1960b), the nature of the occupation-requirement factors which form the basis of occupation rating (Madden, 1960c), the extent to which a rater is familiar with the occupation he is rating (Christal & Madden, 1960; Madden 1960e, 1961), the reliability of occupation-evaluation ratings (Harding & Madden, 1960; Harding, Madden, & Colson, 1960), generalized rater tendencies (Wiley, Harber, & Giorgia, 1959a,

1959b), and methods of rating-scale construction (Madden, 1960d).

Since the rating scale is the vehicle by which a rater makes and communicates his judgments, its importance cannot be over-emphasized. The rating scale may be considered as a measuring device and the task of the rater consists of using this device to make the most accurate measurement he can. His problem is to locate stimuli at the appropriate places on the scale. A study of the intrinsic nature of rating scales, then, appears to be a salient requirement. In addition, such topics as the use of examples to clarify the definition of a rating-scale level, the use of a scaling technique to determine the distance between scale-level definitions, the dimensionality of rating factors, problems of defining rating factors in unequivocal terms, and the format of rating scales are also parts which must be fitted together in an efficient manner if an accurate measuring methodology is to be constructed.

This paper reports a beginning effort to describe the effects upon judgment which are exerted by the format of the rating scale. Format is defined here as the physical arrangement in which the rating-scale definition and levels are presented to the rater for application to stimuli, which in this specific case are occupations.

Many aspects of the construction of a rating scale appear to possess potential determining influences on the judgments made when the rating scale is used. In a previous study (Madden, 1960d) differences in reli-

¹The research reported in this paper was sponsored by the 6570th Personnel Research Laboratory, Aerospace Medical Division, under AFSC Project 773402. This paper is based on Technical Documentary Report PRL-TDR-63-2.

TABLE 1
SCALE LEVELS OF CONDITION B

Scale level	Definition of scale level
+4	Very much more than average
+3	Much more than average
+2	More than average
+1	Slightly more than average
0	About average
-1	Slightly less than average
-2	Less than average
-3	Much less than average
-4	Very much less than average

ability and apparent ease of rating were found depending on whether or not examples were used for the scale-level definitions and whether or not the scale levels were defined at all. When scale-level definitions were not used, reliability was lower than when each level was defined; but examples for each level did not increase reliability. In the present study, interest is centered on whether it makes a difference if the highest scale level is placed at the top or bottom of the list of scale-level definitions; if a graphic device is used to further define the scale levels; if scale levels are numbered from one to n or from +n to -n; or if the level definitions are arranged horizontally or vertically.

METHOD

Rating Scales

Essentially the same rating scale contentwise was constructed in seven different formats. In all seven formats the scale consisted of nine levels, with Level 5 defined as average, with Levels 1 through 4 progressively further below average, and Levels 6 through 9 progressively further above average. When a graphic device was included as a part of the rating scale, it consisted of a bar at each scale level. The percentage of occupations "which normally appear at each scale level" was written in each bar and the length of the bar was proportional to this percentage. The percentages were .04, .07, .12, .17, .20, .17, .12, .07, .04.² A de-

² Sample rating scales have been deposited with the American Documentation Institute. Order Document No. 7823 from ADI Auxiliary Publications Project, Photoduplication Service, Library of Congress, Washington, D. C. 20540. Remit in advance \$1.25 for microfilm or \$1.25 for photocopies and make checks payable to: Chief, Photoduplication Service, Library of Congress.

scription of each rating scale format follows:

Condition A. This rating scale was arranged horizontally with the bars rising vertically. The number of each scale level, 1-9, was printed below each bar; Level 5 was labeled "average"; to the right of Level 5 the word "above" appeared with an arrow extending through Level 9, and to the left of Level 5 the word "below" appeared with an arrow extending through Level 1.

Condition B. In this condition the scale levels were arranged vertically with the bars extending horizontally to the left. The scale levels were numbered and defined to the right of the bars (see Table 1). The verbal scale-level definitions were the same for all remaining conditions.

Condition C. The format of this rating scale was the same as in Condition B except that the bars were omitted.

Condition D. This format varied from Conditions B and C in the numbering of the bars. The lowest level, "Very much less than average," was numbered 1 and placed at the top of the scale. The numbering then increased to 9 at the bottom of the scale, "Very much more than average." Bars were included.

Condition E. Same as Condition D except that the bars were omitted.

Condition F. Same as Condition D except that 9, "Very much more than average," was placed at the top and 1, "Very much less than average," was placed at the bottom. Bars were included.

Condition G. Same as Condition F except that the bars were omitted.

Rating Factors

The raters were 7 groups of 60 basic airmen, 1 group for each condition. (Basic airmen are randomly assigned to flights of 60 or 70 men after arrival at Lackland Military Training Center so that assigning a flight to a condition amounted to randomly assigning airmen to conditions.) The raters were given a sheet containing the rating scale, definitions of 9 rating factors and a list of 15 occupations familiar to everyone. The rating factors were:

Knowledge: Specific knowledge required to perform successfully in the occupation.

Physical Skills: Physical dexterity and muscular co-ordination required to perform the occupation.

Adaptability and Resourcefulness: Degree of versatility, initiative, ingenuity, and judgment used in the occupation.

Responsibility for Money and Materials: Use and misuse or waste and savings of money, materials, and equipment.

Responsibility for Safety of Others: Control of danger involving diligence, effort, and forethought to prevent injury.

Responsibility for Directing Others: Extent to which executive, managerial, and supervisory responsibilities are used.

TABLE 2
SUMMARY OF THE SEVEN CONDITIONS

Condition	Graphic device	Orientation	Highest level	Numberings of levels
A	Yes	Horizontal	Right	1—9
B	Yes	Vertical	Top	+4—-4
C	No	Vertical	Top	+4—-4
D	Yes	Vertical	Bottom	1—9
E	No	Vertical	Bottom	1—9
F	Yes	Vertical	Top	9—1
G	No	Vertical	Top	9—1

Physical Effort: Amount of physical energy required to perform the occupation.

Attention: The level and duration of mental alertness required in the performance of the occupation.

Job Conditions: Environment in which the work must be performed; the discomfort that must be endured.

Table 5 gives the names of the 15 occupations which each subject rated on all nine factors. Subjects were instructed to rate all occupations on one factor before proceeding to the next factor.

RESULTS

Observations from all conditions were converted to the same 1–9 metric. The three-way analysis of variance is summarized in Table 3. All main effects and interaction terms are significant at the .01 level. Estimates of interrater agreement and the reliability coefficient were computed for each factor and condition combination using a method suggested by Lindquist (1953, p. 361). The reliabilities thus computed are in Table 4. Table 5 gives the mean occupation-evaluation score for each occupation and each condition. The oc-

cupation-evaluation score was computed by summing the mean factor ratings for each occupation and condition, then dividing this sum by 9.

DISCUSSION

The significant main effects due to occupations and to factors are to be expected in occupation evaluation. The significant effect due to conditions indicates that the mean ratings assigned to the 15 occupations on the nine factors are not the same across the seven rating conditions, and that the differences are not attributable to chance sampling variation. A finding of this kind is always open to the question of whether or not the statistically significant differences have any practical meaning. The practical implications of a nonchance difference, however, can only be determined by the practitioner.

From Table 5, it is clear that the various conditions tend to result in higher or lower ratings. In Table 6, the seven means for each occupation have been ranked for each

TABLE 3
SUMMARY OF RESULTS OF ANALYSIS OF VARIANCE

Source	SS	df	MS	F	F ₀₁
Occupations	23,380.95	14	1,670.06	620.84	3.00
Factors	10,525.54	8	1,315.69	489.10	4.86
Conditions	611.36	6	101.89	37.87	6.88
O × F	32,063.75	112	286.28	106.42	1.43
O × C	576.59	84	6.86	2.55	1.49
F × C	2,720.71	48	56.68	21.07	1.70
O × F × C	2,787.40	672	4.14	1.53	1.23
Within treatments	150,059.90	55,755	2.69		
Total	222,726.20	56,699			

TABLE 4

RELIABILITY COEFFICIENTS FOR EACH FACTOR AND CONDITION

Factor	Condition						
	A	B	C	D	E	F	G
1	.94	.94	.94	.94	.94	.94	.93
2	.86	.87	.91	.87	.85	.88	.85
3	.86	.86	.82	.83	.84	.79	.89
4	.68	.76	.78	.66	.69	.76	.77
5	.86	.87	.81	.86	.88	.88	.88
6	.86	.90	.85	.85	.86	.90	.88
7	.88	.89	.86	.87	.85	.89	.90
8	.82	.87	.80	.79	.81	.87	.86
9	.72	.24	.44	.38	.65	.66	.62

occupation across the seven conditions. These rankings provide a clear picture of the tendency for some of the conditions to produce low or high evaluations. Table 6 is also a depiction of the Condition \times Occupation interaction. Not only are the occupation-evaluation scores different as a function of the rating condition used, but the condition interacts with the occupation to an extent that is not attributable to chance. Condition C, for instance, ranks 12 of the occupations higher than the other six conditions, 2 of the occupations second, but 1 occupation

TABLE 6

RANKINGS FOR EACH OCCUPATION ON THE SEVEN CONDITIONS

Occupation	Condition						
	A	B	C	D	E	F	G
1	5	3	1	7	2	6	4
2	7	3	1	4	2	6	5
3	5	4	1	6	2	7	3
4	7	6	1	4	2	3	5
5	7	3	2	6	1	5	4
6	2	5	1	7	3	4	6
7	3	6	1	5	2	4	7
8	5	7	1	6	2	3	4
9	6	5	1	7	3	2	4
10	3	5	2	7	1	6	4
11	7	5	1	3	2	6	4
12	6	4	1	3	2	5	7
13	5	6	4	7	1	3	2
14	7	4	1	6	2	3	5
15	7	4	1	5	2	6	3

(aeronautical engineer) is ranked fourth. Condition G results in rankings which range from 2 to 7.

The significant Occupation \times Factor interaction is to be expected in occupation-evaluation, but the Factor \times Condition interaction is not. The statistical significance of this latter term means that the rating assigned

TABLE 5

OCCUPATION EVALUATION SCORES FOR EACH OCCUPATION UNDER EACH CONDITION

Occupation	Condition						
	A	B	C	D	E	F	G
1. Plumber	5.79	5.87	6.22	5.70	5.89	5.72	5.85
2. Librarian	5.07	5.38	5.56	5.34	5.47	5.15	5.21
3. Bank president	6.31	6.35	6.56	6.28	6.44	6.14	6.38
4. Gas station attendant	5.39	5.40	5.93	5.49	5.56	5.51	5.48
5. Police chief	6.70	6.89	6.96	6.83	7.03	6.83	6.84
6. Carpenter	6.53	6.44	6.67	6.33	6.50	6.46	6.44
7. Auto mechanic	6.75	6.53	6.81	6.56	6.75	6.60	6.50
8. Dentist	6.79	6.50	7.02	6.75	6.86	6.84	6.80
9. Truck driver	5.89	5.90	6.08	5.79	6.00	6.02	5.97
10. Electrician	6.78	6.70	6.80	6.60	6.86	6.68	6.73
11. Used car salesman	4.93	5.05	5.47	5.15	5.20	5.00	5.07
12. Short-order cook	4.78	4.99	5.27	5.03	5.22	4.91	4.71
13. Aeronautical engineer	7.07	6.80	7.10	6.68	7.23	7.13	7.20
14. Grade school teacher	6.28	6.52	6.80	6.44	6.71	6.59	6.49
15. Radio announcer	5.06	5.40	5.59	5.29	5.43	5.28	5.42
M	6.01	6.05	6.32	6.02	6.21	6.06	6.07

on the factors depends upon the condition being used. The three-way interaction term, also statistically significant, indicates that the rating assigned to an occupation depends upon the factor-condition combination under which that rating is made.

It is probably appropriate to reiterate at this point that the present study is a probing effort and is not designed to answer the ultimate question of which rating scale format is "best." None of the hypotheses tested can be rejected based upon chance expectancies and it must be concluded that on this basis there are differences in the judgments which are a function of the format of the rating scale. It is suggested that a study designed to select the optimal rating-scale format would require a criterion measure, such as paired comparison scale values, and that this selection should be based upon predictive capability.

REFERENCES

- CHRISTAL, R. E., & MADDEN, J. M. Effect of degree of familiarity in job evaluation. *USAF WADD Personnel Lab. tech. Note*, 1960(Nov.), No. 60-263.
- HARDING, F. D., & MADDEN, J. M. Analysis of some aspects of the Air Force position evaluation system. *USAF WADD Personnel Lab. tech. Note*, 1960(July), No. 60-143.
- HARDING, F. D., MADDEN, J. M., & COLSON, K. Analysis of a job evaluation system. *J. appl. Psychol.*, 1960, 44, 354-357.
- LINDQUIST, E. F. *Design and analysis of experiments in psychology and education*. Boston: Houghton Mifflin, 1953.
- MADDEN, J. M. A review of some literature on judgment with implications for job evaluation. *USAF WADD Personnel Lab. tech. Note*, 1960 (Aug.), No. 60-212. (a)
- MADDEN, J. M. Context effects in job evaluation. *USAF WADD Personnel Lab. tech. Note*, 1960 (Oct.), No. 60-220. (b)
- MADDEN, J. M. A note on the rating of multidimensional factors. *USAF WADD Personnel Lab. tech. Note*, 1960(Oct.), No. 60-258. (c)
- MADDEN, J. M. A comparison of three methods of rating scale construction. *USAF WADD Personnel Lab. tech. Note*, 1960(Nov.), No. 60-262. (d)
- MADDEN, J. M. Familiarity effects in evaluative judgments. *USAF WADD Personnel Lab. tech. Note*, 1960(Nov.), No. 60-261. (e)
- MADDEN, J. M. A further note on the familiarity effect in job evaluation. *USAF ASD Personnel Lab. tech. Note*, 1961(June), No. 61-47.
- WILEY, L., HARBER, H. B., & GIORGIA, JOYCE M. Evidence for a generalized rating tendency. *Engng industr. Psychol.*, 1959, 1, 55-61. (a)
- WILEY, L., HARBER, H. B., & GIORGIA, JOYCE M. Rater tendencies in estimating qualifications required by Air Force tasks. *USAF WADC Personnel Lab. tech. Note*, 1959(Sept.), No. 59-195. (b)

(Received April 22, 1963)

SENSORY FEEDBACK ANALYSIS OF STEREOTELEVISION PURSUIT TRACKING

JOHN D. GOULD¹ AND KARL U. SMITH²

University of Wisconsin

A stereotelevision system, capable of presenting binocular cues for remote depth perception, has been developed for research on problems of optical design and for sensory feedback studies in space science. Preliminary experiments evaluated a color-separation system, which was found to be faulty for research. Detailed visual acuity and stereoscopic acuity tests with a binocular-separation system disclosed that a very adequate and reliable 3-dimensional system can be devised for laboratory studies of remote binocular vision. A specific experiment tested the utility of a nondirectional auditory cue in aiding visual pursuit tracking in depth. Results indicated that the effectiveness of the auditory cue varies as a function of the speed of the target course.

This report demonstrates the application of methods of sensory feedback analysis to the development of the optical design of remote three-dimensional visual-control systems. The main purpose of the experiment was to evaluate a stereotelevision system of our own design for sensory feedback research in visual and space science and to test the hypothesis that nondirectional auditory cues in pursuit tracking will aid performance only at low speeds of target movement.

Research in this field has dealt with both pictorial-panoramic and quantitative or graphic three-dimensional displays, usually in relation to the special problems of space science. Three techniques for pictorial-panoramic display have been developed. Mengle (1958) has described a three-dimensional color-separation system which we have evaluated in this paper. Mauro (1961) has reported an extensive study of the optical accuracy of a three-dimensional color system also. In his system, the images formed by two-lens systems, separated by about 5 inches, are superimposed by means of a beam splitter so that both images fall on the aperture of a single-image orthicon tube. This signal is passed through a filter drum containing Polaroid and color filters "to

effect a color and stereo presentation for direct viewing with special Polaroid viewers." The physics of the system was described and tests of visual acuity reported with one subject. In addition, the binocular system of the sort used here has been tried out in remote-manipulation work.

Morrill and Davis (1961) have done one of the few studies with nonpictorial displays. They had a subject track a target that moved in three dimensions by watching his performance on a two-dimensional dual-beam cathode-ray tube. Elevation information was presented quantitatively on the right channel of the scope while azimuth and range data moved together as the pip moved across the two dimensions of the face of the tube on the other channel. Stereoscopic-radar displays have been devised in which two target sources, corresponding to disparate images of the two eyes, are displayed separately on a tube face. These disparate spots must be viewed with prisms or mirrors to produce fusion and depth. More recently a third type of display has been developed (Bassett and Stone, 1962). In this design, called a volumetric display, since operators view information contained in a clear cylinder, a trail of light spots are displayed inside the cylinder. Here a series of dots are projected from a fixed cathode-ray tube onto a rotating screen inside the clear cylinder. As the screen rotates around the vertical axis of the cylinder, the dots fill out the cylinder in a three-dimensional scene.

¹ Research Fellow, National Institute of Mental Health; now at Thomas J. Watson Research Center, Yorktown Heights, New York.

² This study's funds came from the National Science Foundation (Project 7589) and from the National Institute of Mental Health (Project MH-4469).

Preliminary Research Phase

A first phase of the research investigated a color-separation stereotelevision system that was recommended for handling of nuclear materials (Mengle, 1958). This color chain is composed of two RCA Model TK-201 television cameras, equipped with 2-inch Kodak Ecklar lenses and an RCA 24-inch color kinescope. In this apparatus the depth effect is achieved by means of disparate projection of the red and green components of the color kinescope on the same monitor. One camera is connected to the green color gun, while the other camera feeds into the red gun. The images formed on the television screen are slightly disparate, corresponding to the different angles of incidence of the two cameras with respect to the target.

With this technique, the subject wears a pair of color-filter spectacles. One eye is covered with a green filter and the other eye with a red filter. By wearing these spectacles the subject fuses the two disparate images into one, and views his performance stereoscopically in the television receiver. In order to adapt this system for use in research, the two cameras were mounted on an adjustable, heavy-duty support. The camera supports and main stand were built so that the camera separation could be controlled. The angle of regard of each camera in both the vertical and horizontal planes could be adjusted separately.

This system was evaluated in a series of exploratory studies in which both visual acuity functions and overt visual performances were tested. These studies compared single-chain vision (one camera) with binocular-chain vision (two cameras).

Defects were found in the color-separation system. Subjects had great difficulty in fusing the two disparate images because such a system supplies what may be called focal-plane depth. That is, only the disparate images near the primary focus of the camera are properly spaced and focused for effective vision. Images in the peripheral field in this type of system are too widely separated for fusion.

Visual acuity with this color-separation system was poor. This was due to the fact that the scan-line resolution of a single-color chain of the color kinescope was not as good as the ordinary black-white closed-circuit system.

This poor image, combined with the focal-plane depth effects, made the overall system inadequate for critical research.

Developmental Research

A stereotelevision chain was devised in which the monitor fields of the two cameras were completely separated and fused by means of prisms. This was made to avoid the basic defects of the color system. The apparatus developed is shown in Figure 1. The two TK-201 cameras were connected separately to two matched 8-inch Conrac Monitors (Model CN A8/C). These monitors were mounted in a single rack adjacent to one another. A multiplex arrangement was devised so that the signal from each camera could be interchanged with either one of the two monitors. Mounted on the face of these monitors was a taped viewing hood containing a set of adjustable prisms by which the operator fused the two disparate television images. The prisms in this hood could be adjusted for interpupillary width and for prism power independently in each eye in both the horizontal and vertical meridians. This viewing hood extended 24 inches from the television monitors. With the addition of the prisms the subject's eyes were approximately 25 inches in front of the television screens at all times. The different experiments conducted were designed in terms of the hypothesis that the stereotelevision viewing would aid performance in those situations in which depth factors were of significance.

The general method in these studies was to compare vision with single-camera and two-camera viewing. Since only one camera was used for the control "monocular" condition of viewing, its output signal was fed into both monitors and the subject had to "fuse" the two monitor displays, as in the stereotelevision viewing. This procedure controlled for any favoring of one type of viewing over the other by differential head and body movements. It also eliminated differences in visual acuity between the "single-chain" or "monocular" condition and the stereotelevision condition due to use of the prisms to produce fusion, and it equated the binocular-fusional effects in the stereotelevision and single-camera conditions of viewing.

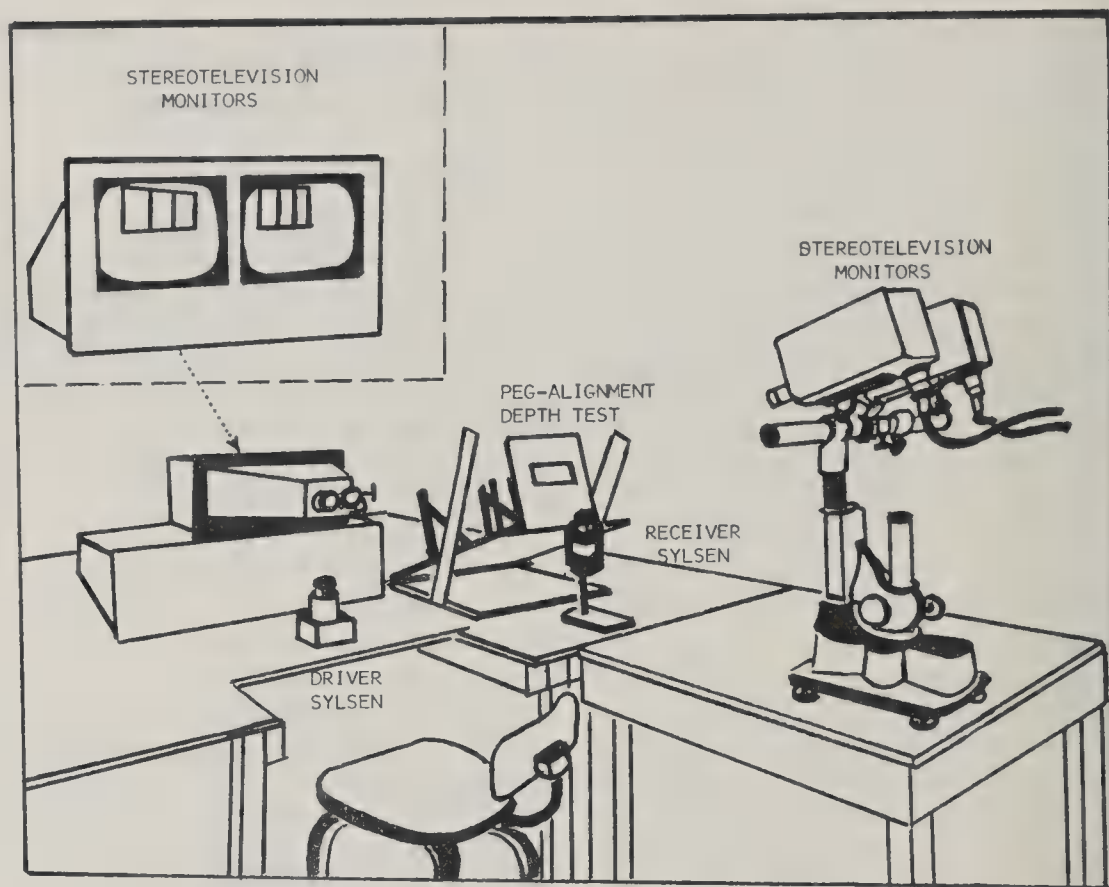


FIG. 1. View of the binocular separation system with the peg test in place.

Stereotelevision Visual Acuity

In making visual acuity tests an arrangement like that shown in Figure 1 was used, except that the performance situation was replaced by a screen on which a visual acuity test chart was projected by a Clayson Acuity Meter. The objective was to obtain standardized measures of visual acuity when the test objects were observed by means of the three-dimensional television arrangement. A 25-character Snellen acuity chart was projected on a high reflectance screen. The acuity meter made it possible to continuously vary the projected size of the letters while always keeping them in focus on the screen. The projector was located 2 meters from the screen on which the letters appeared. The cameras were mounted above the aperture of the Clayson projector. The distance between the lenses of the two television cameras was 23 centimeters.

The letters of the test chart were projected on the screen starting from the smallest size and proceeding to larger projected sizes, i.e., an ascending method of limits, to reduce the

possibility of the subject memorizing the chart.

The Clayson projector gives values of acuity based upon an equal distance from the screen of both the viewer and the projector. For example, if the subject can recognize the 25 letters when both he and the projector are 20 feet from the screen, and the letters subtend an angle of 5 minutes (1 minute between the horizontal bars of an E), then he is said to have 20/20 eyesight or 1.0 decimal acuity. However, appropriate corrections were made in this experiment to compensate for the fact that the stereotelevision system reduced the size of the letter images by a factor of 2.74, thus reducing the acuity, and also for the fact that the subject viewed the fused images 25 inches from the television screens.

The results are presented in terms of the corrected decimal acuity and the minimum angle subtended by each letter (the reciprocal of decimal acuity).

Preliminary study revealed approximately the smallest size of the letters that subjects

TABLE 1
MEASURE OF VISUAL ACUITY IN 11 SUBJECTS USING STEREOTELEVISION VIEWING

Decimal acuity	.39	.43	.52	.60	.69	.78	.91	.95
Minimal angle subtended (min.)	2.56	2.33	1.92	1.67	1.45	1.28	1.10	1.05
Average number of letters recognized correctly	24.2	23.8	23.1	20.6	18.1	12.8	9.19	4.9
Recognition (%)	96.8	95.2	92.4	82.4	72.4	51.2	36.7	19.6

could recognize. Each session started from this point, in order to minimize the possibility of the subject becoming too familiar with the chart. Measurement was made at eight successive increases in the size of the letters, corresponding to decimal acuities on the Clayson apparatus. Room illumination and monitor brightness were adjusted for maximum screen clarity.

Eleven individuals were used in this experiment. Instructions were that a pattern of 25 letters would be presented on the screen and that each one should be recognized. Only one trial was given at each of the eight sizes of the letters. The order in which the subject had to recognize the letters was randomly varied on each trial to further minimize any memorization.

The 11 individuals who served in this experiment returned 1 week later. At this time the same procedure was repeated. This second session was held as a check on the stereotelevision system itself. It was assumed that visual acuity does not change over a week's time; thus, any significant change in the results of the two sessions would be attributed to variations in the system itself. The total number of letters recognized at each of the eight settings by each individual was used to measure visual acuity.

The results for the first session of this experiment are presented in Table 1. The top row of this table shows the decimal acuity at each of the eight sizes of projected letters. The second row gives the visual angle subtended at each of the eight projections. The mean number of letters recognized by all subjects is presented in row three. The percentage of the 25 letters correctly recognized is presented in

the bottom row. It may be seen that over 50% of the letters are recognized at a setting of 1.28 minutes. Likewise, 23 of 25 letters are recognized at a visual angle of 1.92 minutes. This stereotelevision recognition is not as acute as typically found for normal minimum-separable vision (an angle of 1 minute corresponds to 20/20 vision).

In order to determine the effects of prismatic distortion present in the situation, measurements were made with the prisms removed. Under these conditions, the subject viewed only one television monitor. It was found here that 50% of the letters could be recognized at a visual angle of just under 1 minute (.92 minute). Thus, an increase of .36 minute of visual acuity occurs when the prisms were not used. It is possible that some of this increase in performance was the result of practice in recognizing the different letters, since in all cases the condition of the prisms followed that of prismatically fused stereotelevision.

The results of the second week's session were essentially the same as those of the first week. A chi-square test was applied to the two distributions of scores in order to ascertain whether they differed significantly from each other. A nonsignificant chi square of 1.41 was obtained (with 7 degrees of freedom a chi square of 14.07 would be needed to show that the distributions differed significantly from each other). It was concluded that, since the scores of the 2 weeks were so similar to each other, there was little if any variation in the stereotelevision system itself over a week's time. The similarity of the 2 weeks scores would also indicate that if subjects did some memorizing of the chart during the first ses-

sion they did not retain the learning for the second session.

Stereoscopic Acuity

The moving-peg depth test shown in Figure 1 was used to measure relative acuity of depth vision with single-camera and two-camera viewing.

The test used in this study consisted of two identical black pegs, one movable and the other stationary, mounted upright in parallel tracks that were separated by 3.8 centimeters. The movable peg slid along its track toward and away from the front of the test apparatus. The direction of this movement was regulated by the subject, using a system of selsyn motors, while he viewed his performance remotely on stereoscopic television. The subject turned a handknob on the end of the shaft of the drive selsyn, located directly below the viewing hood. This drive selsyn fed its signal into the selsyn receiver motor so arranged that its shaft turned a pulley system which precisely changed the position of the movable peg.

The cameras were spaced 32 centimeters apart. This distance gives a stereoscopic angle considerably larger than the average interocular distance of the subject. The midline distance from the stationary peg to the lenses of the two cameras was 1.8 meters.

Two conditions, stereoscopic presentation and single-camera viewing, were used in the experiment. In both conditions, the subject fused the two pictures with prisms. The subject fixed his head against the anterior headrest of the prisms to avoid head movements and thus cues of motion parallax. Twenty-four individuals served as subjects in both viewing conditions of the experiment. Order of conditions was counterbalanced between subjects. In each condition the subject made five depth adjustments of the movable pegs. The mean of these five judgments was used for data analysis. Performance was measured in millimeter deviations of the movable peg from the standard.

The results of this experiment showed that the 24 subjects averaged 55.1 millimeters deviational error under stereoscopic presentation of the stimuli, while under the single-camera conditions they averaged 78.5 milli-

millimeters deviation, or 23.4 millimeters poorer than under the stereoscopic condition. This difference between the means was found to be statistically significant below the 5% level of chance occurrence ($F = 5.02$; $df = 1/46$). Regarding the angle of stereoscopic visual acuity subtended at these two error values, measurements are based upon the angle subtended by the television cameras and the absolute deviation of the adjustable peg from the standards. These measurements are corrected for the demagnification factor (.3649) of the television system and the subject's distance from the television monitor. Thus, a stereoscopic acuity angle of 16 minutes 24 seconds was found for the two cameras and one of 23 minutes 31 seconds, found for the one-camera condition. These values are considerably larger than those obtained with direct viewing.

Effect of Variation of Camera Separation

The depth-perception tests just described were extended to the study of the effects of varying the separation between the two cameras. Increasing the distance between the optical axis of the two cameras is analogous to increasing the interocular distance in the individual inasmuch as the cameras act as substitute eyes giving disparate images of a certain magnitude of binocular parallax. Also, such increases generally lead to additional strain in maintaining fusion (Mauro, 1961).

In this study, the midline distance from the standard peg to the cameras was 125 centimeters. Seven different interaxial distances between the two cameras were used. These were 22, 25, 30, 35, 40, 45, and 50 centimeters. An experimental design was employed that permitted randomization of the order of these conditions. Accordingly, seven subjects were used and each received a random order of presentation of the distances. The only restriction was that each interaxial distance had to be in each ordinal position (OPS) once and only once.

Subjects were instructed to adjust the movable peg until it appeared in the same plane as the stationary peg. Two practice trials were run, and experimental observations were then begun. During these, the subject received no information as to his

accuracy other than the stereotelevision information. Five trials were run at each interaxial distance for each individual. The mean of these five trials was used for data analysis.

The data of this experiment along with that obtained with two superior subjects is shown in Figure 2. The top curve in this figure gives the data for the study just described using the seven interaxial distances between 22 and 50 centimeters. Except for the values obtained at a separation of 50 centimeters the mean error decreased systematically as the interaxial distance was increased.

Because it was judged that the increase in mean error at a separation of 50 centimeters was an artifact produced by variation of the background display as the cameras were separated by this wide value, the observations were repeated with two female subjects whose performance in this task was known to be superior. With these subjects,

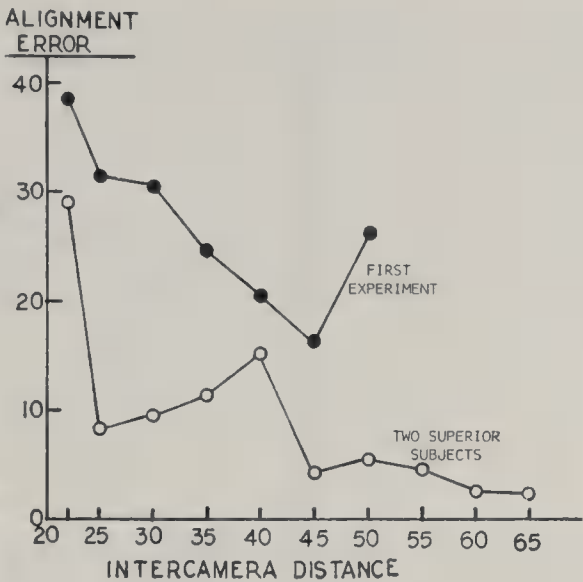


FIG. 2. Curves showing variation in stereoscopic acuity with different camera separation settings.

the background variation was eliminated and the magnitude of camera separations was increased over those used in the first study.

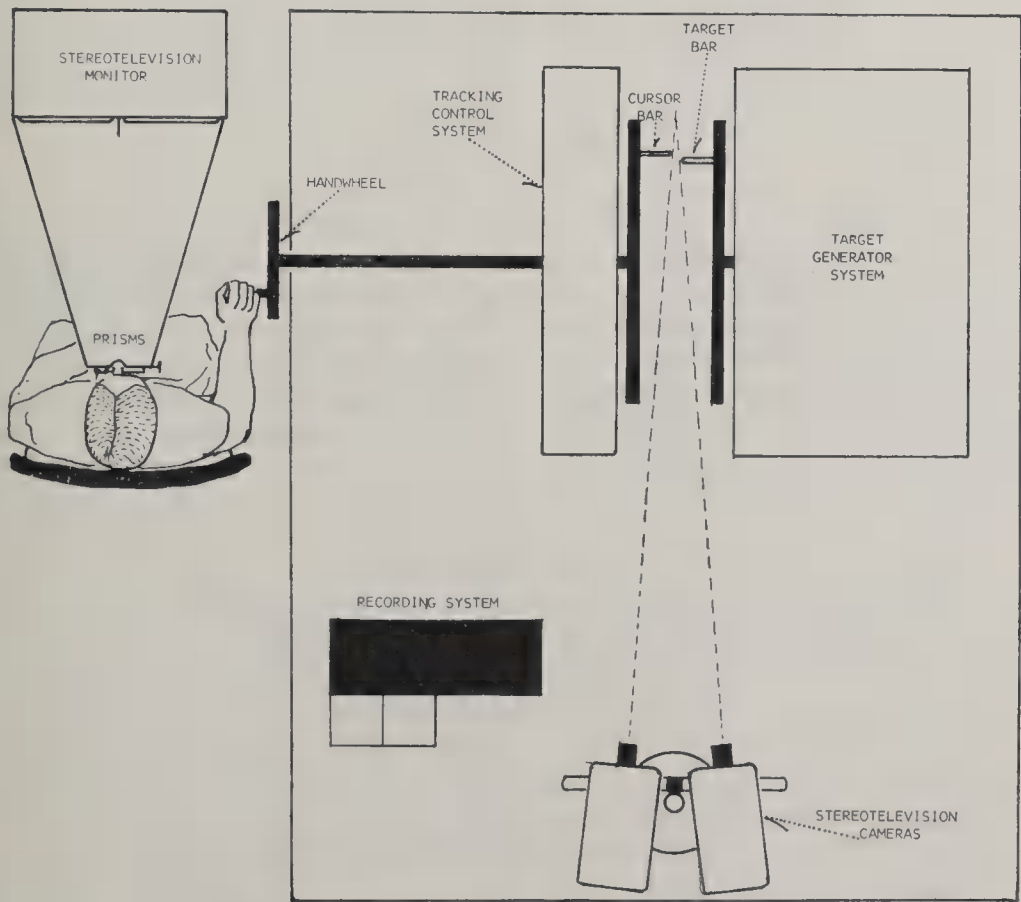


FIG. 3. Overhead view of stereotelevision tracking system. (The camera monitor connections are multiplex.)

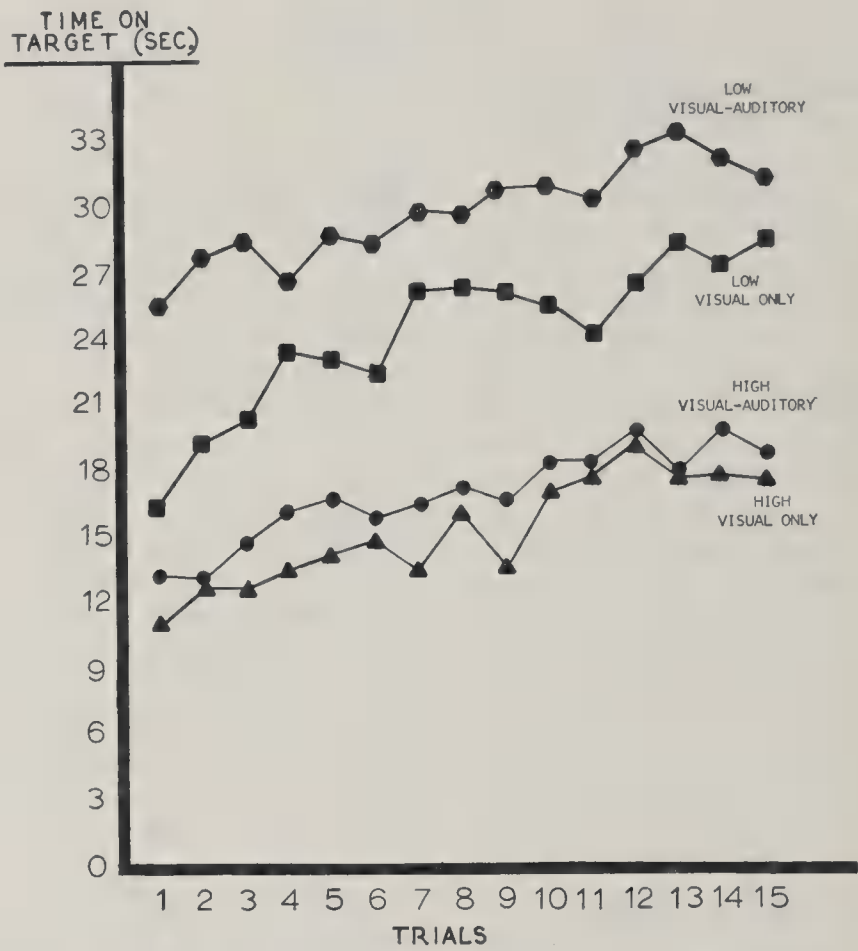


FIG. 4. Learning curves of time on target for different tracking conditions.

Under these conditions, the mean error continued to decrease as the camera separation was increased to 65 centimeters. The data given for these two subjects is based on five settings at each of the 11 interaxial distances.

Table 2 shows the stereoscopic acuity angle subtended at the eyes of the subject for each of the absolute-error values in Figure 2. The

calculations in Table 2 are based upon the interaxial camera distance, the distance from the cameras to the stationary peg, the stereo-system demagnification factor, the distance of the subject's eyes from the television monitor, and the interocular distance. In general, the results show that although the angular value remains approximately the

TABLE 2
STEREOSCOPIC ANGLE SUBTENDED IN DEPTH JUDGMENTS CORRESPONDING TO THE ABSOLUTE ERROR AS SHOWN IN FIGURE 4

Criterion	Intercamera distance (centimeters)									
	22	25	30	35	40	45	50	(55)	(60)	(65)
All subjects	10 min. 45 sec.	10 min. 45 sec.	12 min. 31 sec.	9 min. 36 sec.	12 min. 12 sec.	10 min. 33 sec.	17 min. 4 sec.			
Two superior subjects	8 min.	2 min. 45 sec.	4 min.	5 min. 33 sec.	8 min. 39 sec.	2 min. 46 sec.	4 min. 8 sec.	3 min. 56 sec.	2 min. 21 sec.	2 min. 21 sec.

same as the cameras are moved further apart, the absolute error decreases. This naturally has implications for the design of stereo-remote-control systems. In addition, zoom and telescopic lenses may easily be mounted on this two-camera system. These lenses will provide the subject with greater potential in making depth judgments of targets much farther away than used in the present experiment.

Auditory Aid in Depth Tracking

In this study subjects used a remote-control stereoscopic television-tracking system, shown in Figure 3, in order to perform a direct-pursuit tracking task. The target moved in a course generated in depth. The target was a 4-inch peg, $\frac{1}{2}$ inch diameter, mounted perpendicularly near the rim of an 18-inch disk. This disk was rotated at variable speeds by a cam shaft that allowed the target to move in a circular fashion toward and away the front of the display, with nine reversals in direction during each 1-minute trial. The subject's cursor was an identically mounted 4-inch peg on a similar 18-inch disk facing the target. Each peg rotated independently around the center of its respective disk.

Two target speeds and two modes of feedback were used in this experiment on remote tracking. The mean rate of the high-speed target course was 16.63° per second and the mean rate of the low-speed course was 6.79° per second. That is, the high-speed course on the average was 2.67 times as fast as the low-speed course. In one feedback condition, the subject used stereotelevision only to control his cursor. In the second feedback condition, the visual-auditory, in addition to viewing his performance under stereotelevision the subject heard a tone whenever he had the target and cursor aligned. Nine subjects completed fifteen 1-minute trials under each of the four experimental conditions over a period of 5 days while tracking remotely by means of stereoscopic television. A photocell circuit was located in the tips of the target and cursor to allow the measurement of time on target.

The results of this experiment are shown in the learning curves plotted in Figure 4. It

may be seen that subjects were significantly more accurate ($p < .001$) on the low-speed target course, where they soon learned to stay on target 50% of the time. Performance increased with practice in all four conditions ($p < .001$). In addition, while tracking behavior was significantly better ($p < .001$) under the visual-auditory condition, further analysis showed that only with the low-speed course did the addition of auditory feedback significantly improve performance. Thus a significant interaction ($p < .001$) between speed of the target and type of feedback occurred.

The results show that only on relatively slow-speed target courses is tracking behavior aided by the addition of auditory feedback of performance. In slow-speed tracking, where precise positioning movements are dominant, the subject could—and did—use the non-directional auditory signal in positioning his cursor around a limited area of the target course until he heard that he was, in fact, on target. With the high-speed course, the use of a specific, all-or-none "on-target" signal is less significant because the faster the target moves, the more the operator must predict its course and adjust his cursor accordingly. Hence, the utility of the on-target auditory signal decreases as a function of target speed.

CONCLUSIONS

1. Instrumentation for a stereotelevision system was adopted after extensive testing and evaluation of a two-color and a binocular-separation system. The former was found to be inadequate for accurate visual research; however, the binocular-separation system described here was fully developed for further experimentation.

2. In devising a binocular-separation system, two television cameras were mounted on the same stand and the output of each of these was fed into an adjacent monitor. A viewing hood was placed on the front of the two matched monitors. The subject fused the two disparate televised pictures by means of prisms.

3. A series of initial experiments was conducted to determine the relative efficiency of visual performance with three-

dimensional television. It was found that stereotelevision leads to significantly better visual performance than the two-dimensional television display of visual feedback of motion. By using this binocular-separation method of stereotelevision, it was also found that absolute errors in the adjustment of visual stimuli in depth can be made to approximate the effectiveness of ordinary binocular vision by increasing the difference between the optical axes of the two-camera sources. However, the angle of stereoscopic visual acuity formed with this television system is poorer than that with normal vision.

4. A special study demonstrated that a nondirectional auditory cue aids visual depth tracking only with relatively slow speeds of target movement. Accordingly, the function

of an auditory aid in tracking can be comprehended in terms of its sensory-feedback properties as contrasted to its reinforcement characteristics for learning.

REFERENCES

- BASSETT, R. C., & STONE, J. T. Concepts and requirements for volumetric three-dimensional displays. In Institute of Radio Engineers, *International Congress on Human Factors in Electronics*. Chicago, Ill.: IRE, 1962.
- MAURO, J. A. Three-dimensional color television system for remote handling operation. *USAF ASD tech. Rep.*, 1961, No. 61-430, 103-168.
- MENGLE, L. I. Three-dimensional TV system. *Radio TV News*, 1958.
- MORRILL, C. S., & DAVIES, B. L. Target tracking and acquisition in three dimensions using a two-dimensional display surface. *J. appl. Psychol.*, 1961, **45**, 214-221.

(Received April 29, 1963)

SOME DETERMINANTS OF JOB SATISFACTION: A STUDY OF THE GENERALITY OF HERZBERG'S THEORY¹

ROBERT B. EWEN

University of Illinois

This paper criticized the Herzberg theory that certain work-situation variables ("satisfiers") produce positive, but not negative, job attitudes, while other variables ("dissatisfiers") produce negative, but not positive, job attitudes. Several deficiencies in the methodology of the Herzberg study were discussed. These were: the narrow range of jobs investigated, the use of only 1 measure of job attitudes, the absence of any validity and reliability data, and the absence of any measure of overall job satisfaction. It was concluded that generalizing the Herzberg results beyond the situation in which they were obtained is not warranted.

A recent study by Herzberg, Mausner, and Snyderman (1959) found that the determinants of job satisfaction ("satisfiers") were qualitatively different from the determinants of job dissatisfaction ("dissatisfiers"). It was stated that:

the three factors of work itself, responsibility, and advancement stand out strongly as the major factors involved in producing high job attitudes. Their role in producing poor job attitudes is by contrast extremely small. Contrariwise, company policy and administration, supervision (both technical and interpersonal relationships), and working conditions represent the major job dissatisfiers with little potency to affect job attitudes in a positive direction. . . . Poor working conditions, bad company policies and administration, and bad supervision will lead to job dissatisfaction. Good company policies, good administration, good supervision, and good working conditions will not lead to positive job attitudes. In opposition to this . . . recognition, achievement, interesting work, responsibility, and advancement all lead to positive job attitudes. Their absence will much less frequently lead to job dissatisfaction [pp. 81-82].

It was also stated that:

It would seem that as an affector of job attitudes salary has more potency as a job dissatisfier than as a job satisfier [p. 82].

These findings are in direct opposition to the traditional idea that a given variable in

the work situation can cause both job satisfaction and job dissatisfaction.

It is in general difficult to compare the results of other research to the findings of Herzberg et al. (1959). For example, supervision is a dissatisfier in the Herzberg schema. However, the supervisor may be a source of recognition, which is satisfying. Similarly, salary is a dissatisfier, but it may represent achievement and recognition, which are satisfiers. Such distinctions have usually not been made in other studies. Therefore, this discussion will concentrate primarily on the Herzberg study.

Various procedures of questionable merit were used in the Herzberg study. Some of these may have been responsible for the singular results that were obtained.

Narrow Range of Jobs Investigated. The Herzberg study investigated only engineers and accountants, which represents only a small sample of the jobs which might have been studied. Data obtained from nine different locations were combined into one analysis, and consequently the effect of different situations could not be ascertained. Herzberg et al. (1959) attempted to demonstrate the generality of their findings by tracing the history of human work and showing that they could account for various historical phenomena in terms of their theory. This does not constitute an adequate test of the generality of the theory. It is necessary to replicate the findings with different workers in different job situations, but the authors did not do this. Instead, sug-

¹ This report is based upon a thesis submitted in partial fulfillment of the requirements for the AM degree at the University of Illinois, June 1963. A modified version of this paper was presented at the annual meeting of the American Psychological Association, September 1963.

The writer is indebted to Harry C. Triandis, who directed the research, and to Charles L. Hulin, for suggestions and criticism.

gestions and prescriptions were made for industry on the basis of the one study (Herzberg et al., 1959, pp. 120 ff.).

Use of Only One Measure of Job Attitudes. The Herzberg study used only a semistructured interview to measure job attitudes. This would be acceptable if the study were only of an exploratory nature. In view of the high generality ascribed by the authors to the results, however, the single method of measurement raises questions as to the generality and validity of the findings. The problems of attitude measurement have not as yet been completely resolved and no one method has been shown to be adequate. The need for more than one method of measurement has been effectively argued by Campbell and Fiske (1959).

It is possible that some or all of the Herzberg results were due to the method of measurement that was used. The method was a critical incidents technique; subjects (Ss) told of times when they were particularly happy (or unhappy) and described the cause of their feelings. This procedure could have led to biased results. For example, achievement and advancement were found to be satisfiers. It is likely that when these variables are causes of satisfaction, a critical incident will occur (the employee finishes a difficult job or he is promoted). However, it is difficult to see what incidents would accompany no achievement, or not being promoted. Hence, the critical incidents technique would make it appear as if these variables caused only satisfaction, since only then would a critical incident occur. This is of course only speculation, but the possibility of bias due to the method of measurement employed cannot be discounted when only one method is used.

No Validity and Reliability Data. Herzberg et al. (1959) presented no evidence for the validity of the semistructured interview used in their study. No parallel-form or test-retest reliability coefficients were reported.

No Measure of Overall Satisfaction. Inasmuch as the Herzberg study claimed that satisfiers caused job satisfaction and dissatisfiers caused job dissatisfaction, it would seem desirable to have included measures of overall job satisfaction in the study, but this was not done. Thus, there is no basis for assuming that the factors described in the crit-

ical incidents caused overall job satisfaction (or dissatisfaction). Smith and Kendall (1963) have shown that a worker may dislike some aspects of his job yet still think that it is acceptable overall because "as jobs go, this isn't bad." Similarly, workers may dislike the job despite many desirable characteristics. Likert (1961), questioning the merit of rank-order studies of the determinants of job satisfaction, made a similar point:

if employees say that the thing they like best about their job situation is the clean, well-lighted space in which to work, it does not follow that this factor is most important in producing favorable over-all attitudes. It is even possible that those who give this as their first choice have the least favorable over-all attitudes toward the company. Similarly, the items which are reacted to least favorably cannot be interpreted as the variables which are most important in producing the unfavorable over-all attitudes. . . . It is not the level of favorableness or unfavorableness of response to an item which shows the importance of that item in influencing the over-all job attitudes. Its importance is revealed by the extent to which it is correlated with the total or over-all job attitude score [p. 195].

It is evident, then, that Herzberg et al. have made statements about the causes of overall job satisfaction and dissatisfaction without having any data relevant to overall job satisfaction and dissatisfaction on which to base the conclusions.

The recommendations and generalizations made by Herzberg and associates are unjustified in view of the limitations described above. In fact, the authors have disregarded their own statement about the need for further research in new situations (Herzberg et al., 1959, p. 102).

The Herzberg procedure may be contrasted with that of the Cornell Studies of Job Satisfaction. In this investigation, the researchers used the multitrait-multimethod technique of Campbell and Fiske (1959), and took the trouble to determine the convergent and discriminant validity of their four measures of five aspects of job satisfaction (Macaulay, Smith, Locke, Kendall, & Hulin, 1963) as well as obtaining other estimates of validity (Kendall, Smith, Hulin, & Locke, 1963). In the introduction to the Cornell studies, Smith (1963) stated that

Since generality across a wide range of situations was of primary importance, each step in the construction and validation of the scales was undertaken

in a new situation, as different as possible from those used in preceding steps. Thus . . . mean annual earnings for men in the different plants ranged from \$3,080 to \$8,300, locations ranged from Massachusetts to Tennessee, size of plant varied from 75 to 4,000 employees, and percentage of female employees was from 0 to 50%, with some non-union plants and some closed shops. Individual levels of education ranged from 0 to 20 years, and jobs from janitor to top management [p. 12].

The present writer conducted an exploratory study in an attempt to determine the generality of the Herzberg theory. Responses of 1,021 full-time life insurance agents to a 58-item four-point anonymous attitude scale were obtained.² The Ss were divided into two groups, one of 541 Ss who answered in 1962 and one of 480 Ss who answered in 1960. The 1960 group served as a cross-validation sample. For the 1962 sample, the data were factor analyzed by the method of principal components (unities in the diagonal) and rotated by the varimax method (cf. Harman, 1960). Six clearly interpretable factors emerged: Manager Interest in Agents, Company Training Policies, and Salary (dissatisfiers); The Work Itself and Prestige or Recognition (satisfiers); and General Morale and Satisfaction.

The following analysis was conducted for each satisfier and dissatisfier. The Ss were divided into subgroups which were satisfied, neutral, or dissatisfied with respect to the satisfier or dissatisfier in question. The neutral group consisted of those Ss who checked either of the two middle points of the four-point scale. The general satisfaction of the satisfied and dissatisfied groups was compared to the general satisfaction of the neutral group by using *t* tests of significance. The attribute not tested was held constant; e.g., if the effect of a satisfier was being investigated, only Ss who were neutral on the dissatisfiers were used.

The results indicated that Manager Interest in Agents and Training, supposedly dissatisfiers, actually acted like satisfiers in both groups. Salary also acted like a satisfier in the 1960 group; in the 1962 group, salary caused both job satisfaction and job dissatisfaction. In both samples, the Work Itself was a satisfier as the Herzberg theory predicted;

but Prestige or Recognition caused both satisfaction and dissatisfaction. This study, however, has some of the same deficiencies as the Herzberg study, and is hardly conclusive. The same is true for a recent study by Schwartz, Jenusaitis, and Stark (1963), who used essentially the same methods that were used in the Herzberg study and obtained results which supported most of the Herzberg results.

A more extensive research design is necessary in order to adequately test the Herzberg theory. For the present, however, it must be concluded that the nature of satisfiers and dissatisfiers (if such variables do in fact exist) is as yet far from clear, and may be different in different jobs. Further research is necessary in different occupational situations before any definite statements about the problem are made. There is as yet no justification for generalizing the Herzberg results beyond the situation in which they were obtained.

REFERENCES

- CAMPBELL, D. T., & FISKE, D. W. Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychol. Bull.*, 1959, **56**, 81-105.
- HARMAN, H. H. *Modern factor analysis*. Chicago: Univer. Chicago Press, 1960.
- HERZBERG, F., MAUSNER, B., & SNYDERMAN, BARBARA B. *The motivation to work*. New York: Wiley, 1959.
- KENDALL, L. M., SMITH, PATRICIA C., HULIN, C. L., & LOCKE, E. A. Cornell studies of job satisfaction: IV. The relative validity of the job descriptive index and other methods of measurement of job satisfaction. Ithaca: Cornell University, 1963. (Mimeo)
- LIKERT, R. *New patterns of management*. New York: McGraw-Hill, 1961.
- MACAULAY, D. ANNE, SMITH, PATRICIA C., LOCKE, E. A., KENDALL, L. M., & HULIN, C. L. Cornell studies of job satisfaction: III. Convergent and discriminant validity for measures of job satisfaction by rating scales. Ithaca: Cornell University, 1963. (Mimeo)
- SCHWARTZ, M. M., JENUSAITIS, E., & STARK, H. Motivational factors among supervisors in the utility industry. *Personnel Psychol.*, 1963, **16**, 45-53.
- SMITH, PATRICIA C. Cornell studies of job satisfaction: I. Strategy for the development of a general theory of job satisfaction. Ithaca: Cornell University, 1963. (Mimeo)
- SMITH, PATRICIA C., & KENDALL, L. M. Cornell studies of job satisfaction: VI. Implications for the future. Ithaca: Cornell University, 1963. (Mimeo)

² The writer wishes to thank Robert C. Nuckols of the Life Insurance Agency Management Association of Hartford, Connecticut, who provided the data.

PREDICTING SUCCESS IN BUSINESS¹

FRANK J. WILLIAMS

AND

THOMAS W. HARRELL

San Francisco State College

Graduate School of Business, Stanford University²

Correlations were computed between earnings and each of 15 variables for 196 businessmen who had received the Master of Business Administration degree 15 years earlier. 4 correlations were significant at the 5% level. The highest, .24, was for offices held as an undergraduate. The other 3 were grades in elective graduate courses, Masculinity of the SVIB, and undergraduate professors' ratings. The group was reduced by omitting owner-operators. The remaining group constituted 116 employees. The only variable which correlated significantly with an Administrative-Level criterion was the SVIB scale for Personnel Director.

Although colleges continue to be important sources of future business leaders, little is known about how to recognize, at the graduate school admission stage, those individuals who will distinguish themselves later in business career performance. The relative importance of such factors as the undergraduate academic record, extracurricular activities, faculty recommendations, scores on tests predictive of business aptitude, and the like has not been established. As applications for graduate study expand, however, selection criteria become increasingly important.

This problem is also of great concern to employers, for whom selection errors are costly. More generally, the question of locating sources of, identifying, and selecting men who will manage well is of fundamental importance to the well-being of American business; yet the difficulties connected with predicting success are more formidable in business than in any other area. Thus if predictors which are significantly correlated with success can be identified, selection practices of both colleges and industry can be improved, with the end result that on-the-job achievement will be greater.

METHOD

The present research is a predictive study, whose purpose is to determine which, if any, of a number

of factors are related to business success. The possible predictors considered were those which are available to employers at the time of a student's graduation. They include undergraduate and graduate grade-point averages, scores on selected scales of the Strong Vocational Interest Blank, intelligence, faculty ratings, and extracurricular activities. The population studied consisted of a group of 196 male MBA graduates of the Graduate School of Business, Stanford University, classes of 1927 through 1943, who live in the six San Francisco Bay Area Counties of Marin, San Francisco, San Mateo, Santa Clara, Contra Costa, and Alameda. This represents a relatively homogeneous geographical region in which there is uniformity of living costs and intraindustry salary rates. Graduates employed in such nonbusiness activities as education, social and governmental services, and the like were not included in the population.

The 196 individuals were studied in two ways. First the entire group was studied, using salary as a criterion of success. Then the owner-operators were eliminated from the larger group, leaving for study an employee group of 116 men whose success was measured in terms of a criterion composed of organizational level of authority, degree of participation in guiding company-wide policies, and salary.

Criteria

In this study, the source of the criterion information was a set of questionnaires collected in a 1958 Alumni Survey. From this information two criteria of business success were derived, one for the total group of 196 Stanford MBA graduates, and the other for the employee group of 116 men selected from the total group.

If the criterion to be used is success in business performance, then salary seems to be one good way of measuring it. This criterion, adjusted for the length of time that the subjects (Ss) had been out of graduate school, was used for the total group of 196 graduates. The purpose of the adjustment was to give a higher degree of success to an individual who had achieved a certain income in a shorter time than to one who had achieved that same income in a longer time. These income figures were found to be

¹ This is a presentation of part of a PhD dissertation by Williams on the same subject. The faculty committee was Quinn McNemar, W. A. Spurr, and T. W. Harrell, Chairman.

² Appreciation goes to Fred Schuster, Marianne Gowen, and Lucy Burnham for their editorial assistance.

approximately linear when plotted on ratio paper and a least-squares straight line was fitted to the logarithms of the reported incomes, omitting the indeterminate ones at the upper (above \$50,000) and the lower (below \$5,000) ends of the scale. Lines were then drawn at half standard error distances from the least-squares line, separating the individuals into 12 groups. Those individuals whose incomes were more than 2.5 standard errors below the center line were assigned a criterion score of 1, those between 2.0 and 2.5 standard errors below were scored 2, and so on to the highest score of 12. The distribution of criterion scores was closely fit by a normal curve, there being only a moderate amount of skewness in the direction of the larger scores. Median and modal scores were both 6, and the mean score was 6.6. Proportions of individuals who fell above and below average success for various groupings by years were not significantly different.

An employee subgroup from the original population was formed by first eliminating all owner-partners from the total group. Some of the men in the remaining group were then eliminated from the study because their reported vocational histories were either incomplete or inconsistent in certain respects. Finally the "inheritors"—men known to be connected either by blood or marriage to high corporate officials, or who reportedly got their job through family connections—were eliminated. There remained a group of 116 male employees of predominantly middle-size to large corporations, 78% being employed in companies of more than 1,000 employees, and 43% in companies of more than 15,000 employees. Typically the group may be described as a middle-management group whose median level of authority is three levels below the chief executive and whose median annual salary is \$14,500.

For this employee subgroup, the criterion of business success consisted of a composite of three administrative-level indicators from the Stanford Alumni Survey: the level of authority in relation to the president or chief executive, the degree of participation in deciding overall company policies, and salary. The level of authority was determined by *S*'s answer to the question, "At what level of authority are you from the president or chief executive of the entire company or organization—first, second, third, fourth, fifth, sixth or below?" The degree of participation was determined by the *S*'s answer to the question, "Indicate your highest degree of participation in deciding policies for the entire organization in which you are employed," for which he could check: "I am the final deciding authority," "I make direct recommendations to the deciding authority," or "I do not participate in deciding company-wide policies."

Two possibilities were considered for weighting the scores on these subcriteria, weighting them equally or weighting them according to their average correlation with one another. Of the three subcriteria, salary was found to have a lower average correlation with level of authority and degree of participation than the correlation of these with each other. The differences, however, were insignificant and it was decided to

assign equal weights to the standardized scores to form a composite administrative-level criterion. These scores were then located on a 10-point scale, with an adjustment for time out of school being made, as was done for the salary criterion which was used with the total group. The distribution of criterion scores so obtained was approximately normal, with a mean of 5.6 and a median of 6.

Predictors of Success

Fifteen predictors were chosen from the college records of the 196 men included in the total group. These were: 4-year college grade-point average; grade-point average on required courses in Stanford Graduate School of Business; grade-point average on elective courses in Stanford Graduate School of Business; scores on seven scales of the Strong Vocational Interest Blank for Men—Engineer, Production Manager, Personnel Director, Office Worker, Sales Manager, Lawyer, and Masculinity-Femininity; score on Thorndike (or Ohio) test; score on Faculty Rating Blank; scores on three types of extracurricular activities in 4-year college—athletics, nonathletics, and leadership.

The measure of scholastic achievement in the undergraduate college was the graduation average computed to 1 decimal place on the four-point scale, with the numeric grades corresponding to the letters weighted by the number of credit hours. Then to the grade-point average of non-Stanford graduates was applied a so-called "Stanford Correction" which is an adjustment either upward or downward, based on the comparative grades of students who have attended both Stanford and other colleges. In general, the Stanford MBA program specified 90 units of course work. About half of these units were to be earned in required business courses; the remaining units were to be earned in elective courses offered in the Business School and in other divisions of the University.

Of the 196 men in this study, 126 had taken the Strong Vocational Interest Blank for men. All these tests were taken prior to the completion of graduate work.

For admission purposes applicants to the Graduate School of Business, along with all other entering Stanford University students, were required to take a test of intellectual capacity presumed to be predictive of academic achievement. During the years covered by this study, two tests were used, the Thorndike Intelligence Examination for High School Graduates and the Ohio State University Psychological Test. All scores which now exist had been transmuted to Stanford norms, on which scale the mean score for all Stanford University students was 75, the standard deviation was 15. The group studied here was somewhat superior to this standardization population. The mean score of the 188 men in this study for whom scores were available was 81, which placed the average man in this group intellectually at the bottom of the top third of all entering Stanford students.

TABLE 1
STATISTICS FOR THE PREDICTORS OF
BUSINESS SUCCESS—TOTAL GROUP

Predictors	N	Range	M	SD
GPA 4-year college	196	1.9–3.7	2.61	0.45
GPA required courses, MBA	196	1.3–3.8	2.68	0.58
GPA elective courses, MBA	196	2.0–3.9	3.00	0.48
Strong scales				
Engineer	126	2–56	27.58	13.49
Production Manager	126	12–62	36.87	10.02
Personnel Director	126	9–63	36.20	11.00
Office Worker	126	8–66	43.40	10.39
Salary Manager	126	14–56	37.76	9.04
Lawyer	126	12–59	31.53	9.89
Masculinity-Femininity	126	19–75	48.52	10.34
Thorndike (or Ohio) test	188	50–113	81.13	14.07
Professors' ratings	196	2.3–4.9	4.06	0.42
Activities—athletics	196	0–4	0.85	0.94
Activities—nonathletics	196	0–17	2.98	3.03
Leadership	196	0–10	1.30	1.63

Each applicant for admission to the Stanford Graduate School of Business was requested to ask at least three undergraduate teachers familiar with his personal qualifications to complete adjectival-rating blanks and return the blanks directly to the school. The traits rated were: sociability, initiative, leadership, emotional stability, dependability, and purposefulness. In computing the average rating, quantities one through five were assigned to the modal points on each of the six scales. The rating score used for each *S* was the average score on all completed scales on the three blanks.

In this study extracurricular activities were classified as either athletic or nonathletic and scores on each of these, as well as a "leadership" score, were computed for each of the 196 men. Each *S* was given one point on the athletic score for each sport in which he had lettered and one point for the nonathletic score for each nonathletic campus activity in which he participated. The leadership score was computed by awarding points for each position of leadership in a campus activity.

RESULTS

Product-moment correlation coefficients were computed to determine the relationship between the 15 predictors described and the two criteria—salary for the total group and administrative level for the employee group.

With the exception of the distributions of the three extracurricular activity variables, which were markedly and positively skewed, the predictor distributions were symmetrical. Ranges, means, and standard deviations of

the 15 variables are shown in Table 1. From an inspection of the various scatter diagrams showing the first order regressions, it appears that all of these predictors are approximately linear and that there is no radical departure from homoscedasticity.

Evidently salary is more predictable in the total group than administration level is in the employee group. Table 2 shows the correlations between the fifteen predictors of success and the two success criteria. There were not high degrees of correlation between salary and any of the 15 individual predictors, but four of them were significant: the M-F score on the Strong Test and the ratings by undergraduate professors, both significant at the 5% level; and the grade-point average on elective graduate courses and leadership in the undergraduate college, both significant at

TABLE 2
CORRELATIONS BETWEEN SALARY AND ADMINISTRATIVE
LEVEL AND FIFTEEN PREDICTORS

Predictors	Admini- strative level	
	Salary	
	Total group (<i>N</i> = 196 ^a)	Employee group (<i>N</i> = 116 ^b)
	<i>r</i>	<i>r</i>
GPA 4-year college	.14	.08
GPA required courses, MBA	.13	.16
GPA elective courses, MBA	.22**	.15
Strong scales		
Engineer	.11	–.10
Production Manager	.08	.05
Personnel Director	.09	.24*
Office Worker	–.03	–.01
Sales Manager	–.05	.10
Lawyer	–.04	.07
Masculinity-Femininity	.19*	.06
Thorndike (or Ohio) test	–.01	–.01
Professors' ratings	.18*	.03
Activities—athletic	.13	.15
Activities—nonathletic	.13	–.05
Leadership	.24**	.14

^a Except for the Strong scales where *N* = 126 and the Thorndike (or Ohio) test where *N* = 188.
^b Except for the Strong scales where *N* = 78 and the Thorndike test where *N* = 109.
* Significant at the 5% level.
** Significant at the 1% level.

the 1% level. Only one correlation was significant at the 5% level for predicting the administrative level and that was the correlation with the Personnel Director scores on the Strong Vocational Interest Blank.

For the total group, the correlations between salary and the six occupational scales of the Strong Vocational Interest Blank were not significant, ranging from minus .05 to plus .11. There was, however, a significant positive relationship between success and the score on the Masculinity-Femininity scale, indicating that those individuals with the stronger masculine interests have a somewhat better chance of success.

The correlation of .18 between undergraduate professors' ratings and salary in the total group was significant at the 5% level. These ratings constitute the only appraisal of nonintellectual personality traits available for this study. Evidently for the total group, those individuals who were above average on these personality traits at college age tended also to be somewhat above average in success later on in business.

There was no indication of any relationship between success and participation in either athletic or nonathletic activity in the total group. The most valid predictor of earnings was leadership in college, with a correlation

of .24. Thus it appears that a man's leadership qualities as evidenced on the campus by his election to presidency or vice-presidency in fraternities, student governments, clubs, etc., may provide a valuable clue to his potential for success in business.

DISCUSSION

An interesting finding in this study was that while the grade-point averages for undergraduate courses and for required graduate courses fell short of significant correlations with the success criterion, there was a significant correlation between grades on elective graduate courses and salary. This difference seems to indicate that scholarship would merit further consideration in business-success studies. Attention might well be directed away from the overall grade-point average, however, toward some figure which would concentrate on the elective course area.

The significant result for the correlation between leadership and salary supports the above reasoning by indicating that distinction in college in chosen areas of interest, when these interests involve either academic or administrative responsibilities, is a possible predictor of success in business performance.

(Received May 6, 1963)

A FACTOR ANALYSIS OF RETAIL CREDIT APPLICATION DATA

JAMES H. MYERS

University of Southern California

22 variables relating to 1200 applications for credit in a nationwide finance company were intercorrelated and factor analyzed. Variables included personal-history information, financial condition and background of the applicant, and data relating to the current transaction. The 6 factors which emerged were identified as: Size of Present Transaction; Applicant Stability; Previous Use of Credit; Personal Income; Loan Duration, and Domestic Work Pattern. Variable communalities were extremely low, with only 1 exceeding .50.

While factor analytic techniques have been applied to a wide variety of psychological and sociological problems, their use in analyzing biographical or personal history information has been quite limited. Not a single published study could be found in which intercorrelations among biodata variables had been factor analyzed, in spite of the fact that such variables have long been used in "weighted application blank" selection tools for both job and retail credit applicants.

The present study is an outgrowth of a study (unpublished) to develop a numerical credit scoring system for use in field offices of a nationwide finance company, headquartered in Los Angeles. Such a system is based on the weighted application-blank approach, as illustrated by several published studies (Durand, 1941; McGrath, 1960; Myers & Cordner, 1957; Myers & Forgy, 1963). The present study involves a factor analysis of personal history, financial history, and current transaction variables involved in customers' applications for credit in the company's field offices.

METHOD

The items of information shown below were obtained from each of 1,200 applicants *who were granted credit* in any of the company's more than 40 field offices throughout the country. One half (600) of these had paid their credit obligations in full, the remainder were delinquent in payments and had been charged off as uncollectible. All applications were made during the years 1958 through 1960. Transactions ranged from the purchase of swimming pools and other home improvements to furniture, including pianos. Cases were drawn randomly from

all field offices, so that the resulting sample would be considered a representative cross section of company business. All of the information below was taken from the application form and other related documents which were completed *at the time of application*.

Personal and Financial Variables

The following information was obtained from each applicant:

1. Net amount of credit requested (total price of item purchased less down payment plus carrying charges).
2. Loan duration in months (*as applied for*, as opposed to the number of months actually paid).
3. Age of applicant.
4. Age difference between man and wife (if applicant is married).
5. Number of dependents (excluding wife).
6. Length of time at present address in years.
7. Telephone at present residence?
8. Occupation (self-employed versus work for others).
9. Length of time at present employment, in years.
10. Monthly income (all sources: man, wife, investment income, etc.).
11. Telephone at work?
12. Do both man and wife work? (If applicant is married.)
13. Investment income (from investments or sources other than present employment).
14. Type of bank account (checking, savings, loan etc.).
15. Present indebtedness (total balance on all installment obligations including present purchase but excluding home mortgage).
16. Number of previous credit accounts whose balance at any time exceeded \$200.
17. Previous high credit (highest balance at any time on any previous credit transaction).
18. Number of previous credit extensions by personal-loan-type finance company (these may also be included in number 16 above).
19. Type of residence (own house versus rent house, versus rent apartment).

- 20. Down payment percentage on present purchase.
- 21. Number of credit references given which subsequent investigation revealed were not paid in a satisfactory manner.
- 22. Account status (paid versus delinquent).

Scaling of Items

Since some of the biodata items were attribute in nature (e.g., rent versus own, type of bank account, occupation), original coding of these items from the application form produced "nominal" scales. It seemed reasonable that these could be transformed into an ordinal form by scaling along a continuum of credit desirability; that is, categories reflecting good credit-payment potential would be given larger numerical values, categories reflecting poor potential would be given smaller values. The example shown below will help to illustrate:

Analyses

The resulting 22 ordinally scaled items were inter-correlated, and principal components factors were extracted by the IBM 7090 computer at the Western Data Processing Center, University of California, Los Angeles.¹ Since many of the later factors emerging had very small loadings, it was decided to include in rotations only those factors with at least one loading in excess of .20. The six factors meeting this criterion were rotated by computer using the "varimax" solution developed by Kaiser (1959).

RESULTS

Rotated factor loadings are shown in Table 2.

Factor A reflects size of the present transaction; the higher the amount of purchase, the higher the individual's total present indebtedness (*including* the present purchase). The apparently opposite (positive) loading on present indebtedness is due to the fact that this item was ordinally scaled in the opposite direction, with highest ordinal values assigned to low indebtedness (done in this manner for purposes of another portion of the study). Therefore, the loading on this variable reflects what would be expected: the greater the present purchase, the greater the total present indebtedness. Also, higher credit amounts on the present transaction are associated with greater previous use of credit.

Factor B reflects a general pattern of applicant age and stability. Age is associated with time at both present address and present

TABLE 1

EXAMPLE OF SCALING NONCONTINUOUS ITEMS

Type of bank account	Number of cases		Nom- Or- inal al code value	
	Good	Bad		
None	124	232	0	0
Savings only	62	52	1	1
Checking only	216	188	2	1
Loan only	20	28	3	1
Checking plus savings	114	56	4	2
Other combination	64	44	5	1
Total	600	600		

job, but not to any great extent with other borrower characteristics. The loading on Account Status is in the favorable direction.

Factor C reflects the frequent previous use of installment credit in making purchases of all types. The opposite loading on present indebtedness indicates that frequent previous use of credit is associated with larger amounts of present installment indebtedness, as would be expected.

Factor D is apparently an income factor, although loadings are low. Higher income families seem more often to have telephones and bank accounts and to be more likely to pay in full.

Factor E reflects loan duration, as distinct from size of transaction (Factor A), although there appears to be some relationship with size in this factor at least. Applicants requesting longer payment terms generally give smaller down payments and are more likely to become delinquent.

Factor F may reflect the domestic work pattern, in that families where both man and wife work are more likely to produce a larger monthly income (among the types of families served by this firm, at least). Here again, loadings are low and interpretations must be made with caution.

DISCUSSION

It is interesting to note that two of the factors were concerned with the transaction, while the remaining four reflected various personal history aspects of the applicant. The relatively low loadings for all factors can be explained by the low communalities of the application variables. Excluding the criterion

¹ Using a program developed by the Division of Biostatistics, School of Medicine, University of California, Los Angeles.

TABLE 2
ROTATED FACTOR LOADINGS

Item	A	B	C	D	E	F	h^2
1. Net credit	-.54	.04	-.03	-.06	-.39	-.23	.50
2. Loan duration	-.16	-.01	-.01	-.15	-.58	-.21	.42
3. Age of applicant	.00	-.59	.05	.03	-.16	.07	.38
4. Age difference	.01	.05	.03	.25	.08	-.06	.08
5. Number dependents	-.10	.19	-.17	-.17	.10	-.16	.14
6. Time present address	-.08	-.55	.07	-.04	.02	.06	.32
7. Telephone	-.08	-.22	-.03	.28	-.02	-.21	.18
8. Occupation	-.30	-.11	-.06	.14	.00	.04	.12
9. Time present job	-.06	-.57	-.03	-.10	-.01	-.03	.34
10. Monthly income	-.19	-.08	-.05	.30	.08	-.38	.29
11. Work telephone?	.02	.11	-.11	.35	.03	.01	.15
12. Both work?	.00	.02	-.03	-.02	-.02	-.38	.15
13. Investment income?	-.05	-.06	-.08	.14	-.18	.05	.07
14. Bank account	-.16	-.10	.10	.34	.07	-.08	.17
15. Present indebtedness	.64	-.10	.26	.04	.32	.18	.63
16. Number of credit references over \$200	-.29	-.07	-.59	.04	-.04	-.17	.47
17. Previous high credit	-.40	-.17	-.40	.14	.09	-.08	.38
18. Number of finance company references	.03	.22	-.51	-.11	-.16	-.20	.39
19. Type residence	.06	-.08	.06	-.14	.01	.33	.14
20. Down payment	.00	.07	.08	.10	.42	-.07	.20
21. Number of unsatisfactory credit references	-.04	.07	-.26	.04	-.15	.08	.10
22. Account status	.03	-.46	.15	.28	.40	-.09	.48

(Account Status), only four variables had communalities in excess of .40, while 11 were at or below .20.

One explanation for such low communalities would be the lack of a large number of biodata-type items available from the application form. However, considering the entire realm of retail credit, the application information requested by this company is comparable to that requested by other finance companies and greatly exceeds applications for most retail stores, gasoline company credit cards, etc. In the retail credit field, then, the information in this study is about as much as any company will have to work from.

A more likely explanation of the low original intercorrelations is that of restriction of range on many of the variables, since the present study was concerned only with applicants who had been granted credit. In the company studied, this was only approximately 65% of the total number applying for

credit, so that a considerable amount of screening must have taken place at the field-office level. The effects of this screening upon factor structure is not known at the present time. The method chosen for scaling the data into ordinal form undoubtedly had an effect upon factor structure also.

REFERENCES

- DURAND, D. Risk elements in consumer installment financing. (Study No. 8) New York: National Bureau of Economic Research, 1941.
- KAISER, H. F. Computer program for varimax rotation in factor analysis. *Educ. psychol. Measmt.*, 1959, **19**, 413-420.
- MCGRATH, J. J. Improving credit evaluation with a weighted application blank. *J. appl. Psychol.*, 1960, **44**, 325-328.
- MYERS, J. H., & CORDNER, W. C. Increase credit operation profits. *Credit World*, 1957(Feb.), 12-13.
- MYERS, J. H., & FORGY, E. W. The development of numerical credit evaluation systems. *J. Amer. Statist. Ass.*, 1963, **58**, 799-806.

(Received May 6, 1963)

QUANTIFICATION OF BIOGRAPHICAL DATA FOR PREDICTING VOCATIONAL REHABILITATION SUCCESS

RAYMOND A. EHRLE¹

To devise an instrument based on biographical data to classify applicants for State vocational rehabilitation services in terms of success and to construct expectancy charts to indicate probability of success of future applicants. The Ss were 200 clients closed in fiscal year 1960 as being employed and 200 closed in fiscal year 1960 as being unemployed, as well as 40 in each category closed during fiscal year 1961. 86 items of personal data were obtained for key K₁ and 20 selected items were obtained for key K₂. Results were: variance between criterion subgroups could be maximized for classification, scores could be derived to classify clients, scores could be combined to establish expectancy charts, and K₂ predicted expectancies better than K₁.

The purpose of this study was to devise an instrument, based on personal and biographical information, to classify applicants for state vocational rehabilitation services in terms of vocational success or failure and to then construct expectancy charts to indicate the probability of vocational success for future applicants for rehabilitation services. It was hypothesized that items of personal data contained in Missouri State Vocational Rehabilitation Form R-4 (the basic intake form) could be combined to yield this information by the empirical method of personal-data analysis described by Welch, Stone, and Paterson (1952).

PROCEDURE

Subjects (Ss) were 200 randomly selected clients closed in fiscal year 1960 as being suitably employed and 200 randomly selected clients closed in fiscal year 1960 as being unemployed, as well as a group of 40 in each category closed during the first 6 months of fiscal year 1961.

A construction group ($N = 200$) composed of equal numbers of employed and unemployed clients, a primary (control) group ($N = 200$) composed of equal numbers of employed and unemployed clients, and a secondary (experimental) group ($N = 80$) composed of equal numbers of employed and unemployed clients were established.

All personal data information was extracted from the "survey interview" portion of the R-4 for each S

¹ Assigned to the Division of Counseling and Testing Services, United States Employment Service, United States Department of Labor, Washington, D. C. Views expressed are those of the author and are not necessarily those of the United States Department of Labor. This paper is based upon a doctoral dissertation completed at the University of Missouri in 1961 under the direction of John F. McGowan.

in the construction group and placed on a worksheet for analysis. The method of analysis consisted of the establishment of variables, and intervals within variables, of personal data items which discriminated between the employed and unemployed subgroups of the construction group. Discriminatory intervals and variables which differentiated were given unit weights of 1, 2, and 0 to make up scoring keys. Intervals were empirically determined, based on frequency of responses, and how well they seemed to discriminate between the criterion subgroups.

Unit weighting was accepted in consideration of its ease of use and relative efficiency and the fact that the approximate ratio within a set of weights is considered more important than their numerical values. Eighty-six items were retained and designated as scoring key K₁. An additional scoring key, K₂, using 20 selected items and using weights previously derived from the construction group, was constructed. This key was to be used for comparative purposes.

After scoring each work sheet, by totaling unit weights appropriate to each variable, score distributions were plotted for construction and cross-validation groups. A cutting score was computed for classification purposes, based on the greatest index of differentiation. Cross-validation group Ss were classified in terms of potential vocational success or failure and the significance of the difference between results obtained was computed. Expectancy charts were then prepared on the basis of score distributions.

RESULTS

Table 1 reveals that no significant difference existed between construction and cross-validation group means. This was interpreted to mean that all groups were of the same general population.

Table 2 reveals that the vocationally successful and unsuccessful subgroups, of each group, were shown to have significantly different means, thus confirming the presence of differentiating variance among the voca-

TABLE 1
SIGNIFICANCE OF DIFFERENCES BETWEEN CONSTRUCTION
AND CROSS-VALIDATION GROUP MEANS

Group compared	86 variables		20 variables	
	<i>M</i>	<i>CR</i>	<i>M</i>	<i>CR</i>
Construction and primary cross-validation	101.25		107.30	
	102.00	.27	107.25	.10
Primary and secondary cross-validation	102.00		107.25	
	98.69	1.11	103.85	.81
Construction and secondary cross-validation	101.25		107.30	
	98.69	.81	103.85	.77

Note.—No critical ratios are significant at the .01 level.

tionally successful and unsuccessful groups.
A cutting score of 90 plus was established for the construction group, based on the index of greatest differentiation. This was accomplished by preparing a table showing the percentage of each group scoring within or above each possible score range. Indexes of differentiation for each possible score range were computed by subtracting the percentages for the vocationally unsuccessful from the percentages for the vocationally successful and disregarding the algebraic sign. The score at which the index of differentiation was greatest is the optimum cutting score.
All percentages of classification in the cross-validation groups resulted in an increase in predictive efficiency of 10 to almost

19% over chance. Scoring with the K₂ key resulted in 67.5 to 68.8% correct classification or a 17.5 to 18.8% improvement over chance. Scoring by K₂ predicted somewhat better than K₁ but results overlapped greatly.
Inasmuch as expectancy charts were subsequently prepared, the product-moment correlation coefficient between criterion and cross-validation groups was not computed. However, by using the index of forecasting efficiency, and Table L in Guilford (1936), an estimated product-moment coefficient of between .45 and .60 was obtained between the criterion and cross-validation group. This estimate is based on an improvement in classification accuracy over chance of between 10 and 19%.
A reasonable overall estimate of the coeffi-

TABLE 2
SIGNIFICANCE OF DIFFERENCES BETWEEN SUBGROUP MEANS

Groups	86 variables		20 variables	
	<i>M</i>	<i>CR</i>	<i>M</i>	<i>CR</i>
Construction Successful Unsuccessful	122.05 80.45	21.67*	137.95 76.65	15.21*
Primary cross-validation Successful Unsuccessful	107.65 96.35	4.55*	120.75 92.75	6.01*
Secondary cross-validation Successful Unsuccessful	105.63 91.75	4.59*	113.25 90.37	3.86*

* *p* < .01.

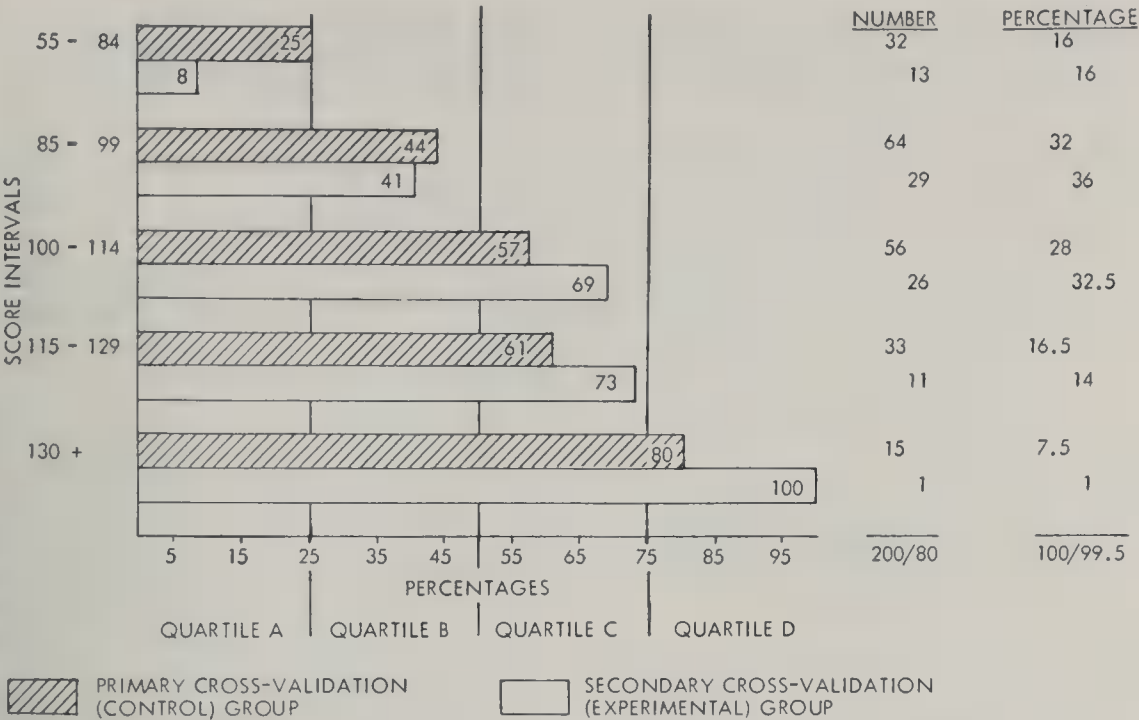


FIG. 1. Expectancy chart based on 86 variables. (Primary cross-validation group results compared with results of the secondary cross-validation group.)

cient of correlation is .53. Such a correlation accounts for 28% of the relationship between the criterion and personal data variables. It also represents a 15% increase in predictive efficiency (1-k) over chance.

Expectancy charts were prepared on the basis of score distributions, taking into account the proportion of those achieving success at different levels. Figure 1 depicts expectancies utilizing 86 variables while

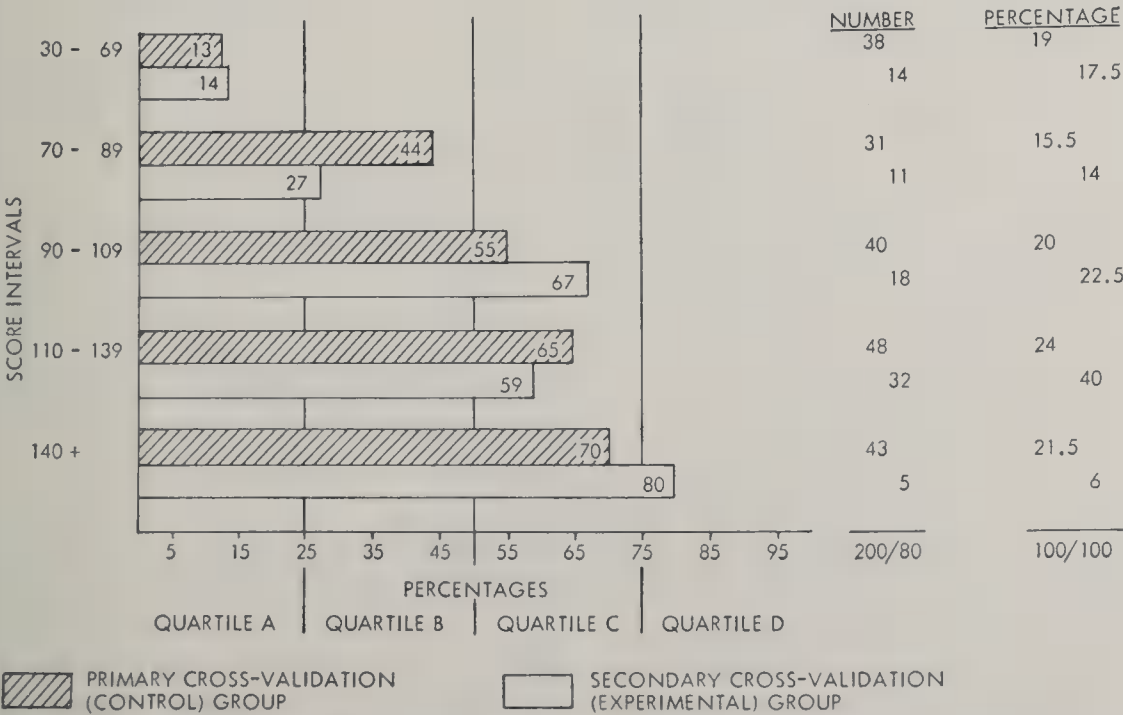


FIG. 2. Expectancy chart based on 20 variables. (Primary cross-validation group results compared with results of the secondary cross-validation group.)

TABLE 3
PERCENTAGES OF CORRECT CLASSIFICATION OBTAINED
USING 86 AND 20 VARIABLES WITH A 90+ CUTTING
SCORE

Group	86 vari- ables (K ₁ Key)	20 vari- ables (K ₂ Key)
Construction (N = 200)	83.5	83.0
Primary cross-validation (N = 200)	60.0	67.5
Secondary cross-validation (N = 80)	65.0	68.8

Figure 2 depicts expectancies utilizing 20 variables.

Individual expectancies in the ranges of 0-25, 25-50, 50-75, 75-100 chances in a hundred that the individual would succeed led to very consistent results.

CONCLUSIONS

1. With these data, it is possible to maximize the variance, in terms of difference scores between criterion subgroups, for classification purposes.
2. Difference scores may be weighted and combined to classify rehabilitation clients into the dichotomous classification of potential vocational success or failure.
3. Scores may be combined to establish expectancy charts to function as employ-

ability indexes for individual predictive purposes. All information required for analysis is available at the time of the initial interview; expectancy charts may be used to determine the rank order in which individuals stand in terms of potential success at the time they are accepted for services.

4. The 20-variable key-predicted expectancies of vocational success better than the 86-variable key and at a much lower administrative and operational cost.

Expectancy charts can be used to statistically estimate an individual's rehabilitation potential, aid in planning individual services, counsel and advise the individual of his probable chances of success, perform many descriptive and comparative analyses, study characteristics of individuals at either end of the scoring distribution, as well as those for whom predictions had gone astray and assist in the performance of various administrative tasks such as budget preparation.

REFERENCES

GUILFORD, J. P. *Psychometric methods*. New York: McGraw-Hill, 1936.
WELCH, JOSEPHINE, STONE, H. C., & PATERSON, D. G. *How to develop a weighted application blank*. Dubuque, Ia.: William C. Brown, 1952.

(Received May 9, 1963)

EFFECTS OF DIFFERENT PATTERNS ON OUTCOMES OF PROBLEM-SOLVING DISCUSSION

JOHN K. BRILHART AND LURENE M. JOCHEM

Department of Speech, Pennsylvania State University

Using an analysis of variance design, some effects of 3 patterns for problem-solving discussion were investigated experimentally. All 3 patterns began with an analysis of the problem, but differed thereafter: ideation-criteria (A), criteria-ideation (B), and solution (C). 135 student Ss were randomly assigned to 27 groups of 5 Ss and a trained leader. All groups discussed 3 problems, using each pattern once. Patterns A and B, incorporating deferred judgment, yielded significantly more ideas ($p < .01$) than Pattern C. There were no significant differences among nights, leaders, or groups. Pattern A yielded significantly more ideas judged to be good than Pattern C. Ss ranked Pattern B significantly lower in order of preference for future use than A or C (χ^2 , $p < .01$). The results tend to discredit patterns recommended in the majority of current textbooks and manuals for discussion leaders and participants.

Ever since Dewey published his essay on *How We Think* (1910), the majority of textbooks and manuals have recommended modifications of Dewey's formulations as patterns for group problem-solving discussion. Most of these patterns closely approximate the following: (a) defining and analyzing the problem; (b) establishing criteria by which to judge proposals; (c) finding possible solutions, or the generating of proposals; (d) evaluating proposals in order to reach a decision; (e) planning how to put the final solution into effect.

There are, however, conflicting views about the desirability of including the discussion of criteria as a separate step. Even among those who include a criteria step, there are differences as to when it should occur (Barnlund & Haiman, 1960, p. 93). A few writers suggest generating solutions before establishing criteria and others suggest discussing values in the immediate context of the solutions being proposed. Osborn (1957) and Parnes (1961) advocate that ideation should precede setting criteria. Combining ideation and evaluation, says Osborn, would be as self-defeating as driving with the brakes on. Studies of individual problem solving reported by Osborn (1962) indicate that ideation before criteria produces more and better ideas than does the criteria-ideation pattern typical of the discussion textbooks.

There has been little empirical study of

patterns for group problem-solving discussion (Golembiewski, 1962, p. 167). Bales and Strodtbeck (1951) found emphasis on orientation, evaluation, and control in subsequent thirds of discussions by small leaderless groups, and Maier established the potential superiority of a detailed pattern over free discussion (Maier & Maier, 1957).

The present study was designed to investigate experimentally some outcomes of three different problem-solving discussion patterns in which ideation, criteria, and evaluation were separated or combined, and in which the sequence of ideation and criteria was changed. Examined were the number of possible solutions proposed, the number of "good" ideas proposed, the quality of final solutions, and subjects' (Ss') satisfaction with final solution, designated leader (L), and the patterns of problem solving.

METHOD

Experimental Design

The design was a three-factor analysis of variance with three patterns, three problems, and nine (Ls). A sample of 135 Ss was drawn from volunteers enrolled in a basic speech course required of all students graduating from the Pennsylvania State University. The Ss were sophomores, juniors, and seniors. On each of 3 nights, 45 Ss were randomly assigned to nine experimental groups of five each.

The nine experimental leaders were selected from upper class and graduate students who had completed

a course entitled "Discussion" taught by the senior author. Five were males, four were females. The *Ls* were given 4 hours of special training, including practice in using the three patterns.

Recorders, volunteers from the senior author's classes, were given brief instruction and randomly assigned to experimental groups.

Problems

Three problems were discussed in the same order by all 27 groups.

Problem 1. The library problem involved a local situation somewhat familiar to all *Ss*:

The Pattee Library has been plagued with books and magazines being mutilated, articles cut out, pages ripped away, and materials being stolen. What might be done to alleviate this problem?

Problem 2. The tourist problem was taken from Taylor, Berry, and Block (1958, p. 28):

Every year a great many American tourists go to visit Europe. But now our country wants to get more European tourists to visit America during their vacations. What steps can you suggest that would get more European tourists to come to this country?

Problem 3. The teacher problem also came from Taylor:

Because of the rapidly increasing birthrate beginning in the 1940s, it is now clear that by 1970 public-school enrollment will be very much greater than it is today. In fact, it has been estimated that if the student-teacher ratio were to be maintained at what it is today, 50 percent of all individuals graduating from college would have to be induced to enter teaching. What different steps might be taken to ensure that schools will continue to provide instruction at least equal in effectiveness to that now provided?

Instructions

After being told the general purpose of the experiment, the assembled *Ss* were given code numbers and group assignments. In order to motivate cooperative intragroup behavior, *Ss* were promised that the group whose three solutions received the highest average rank by expert judges would be awarded \$15. The *Ss* were requested not to discuss the experiment with anyone until after the third night of discussions. All *Ss* on the second and third nights testified that they had not heard either the patterns or problems. Following the request, *Ss* were sent to nearby classrooms in which a circle of six chairs was arranged adjacent to the blackboard, with *L* sitting next to the blackboard and the recorder sitting outside the circle and opposite *L*.

The *Ss* were greeted by *L*, seated, and supplied with name cards which showed first name and code number. After introducing himself and the recorder, *L* read a brief description of the roles of *L*, *S*,

and recorder. He next said, "Let's win that money, be the best group in the study. Okay?" and read the following:

I am going to read a problem to you. Then we will have up to 35 minutes in which to discuss the problem and decide what we believe to be the best solution for it. My job is just to keep the discussion organized and following a particular procedure. It is important that we follow that procedure closely while discussing the problem and working toward a solution. I will not do any of the discussing, propose any ideas, give any hints, nor do any evaluating; I'll just present the problem, explain the procedure, and see that we follow the agenda in order to arrive at a decision.

Next, *L* read an abbreviated outline of the problem-solving pattern, answered questions, read the problem, and guided the group through the scheduled pattern. As soon as the discussion was finished, *L* distributed a brief questionnaire, collected completed questionnaires, and proceeded to the second discussion. This general procedure was repeated for all three discussions, changing only the problem and pattern.

Following the third discussion, *L* thanked the *Ss*, told them how to sign up for a report of the findings of the study, reminded them not to discuss their experience with anyone, and wished all a good night.

Problem-Solving Patterns

Although the sequence of problems remained the same on all 3 nights, the sequence of patterns was systematically rotated in order to counteract any sequential effects. Both Patterns A and B incorporate the "rules" for brainstorming, which *Ls* were coached to administer quite stringently.

When using Pattern A, *ideas-criteria*, *L* followed guidelines laid down by Osborn (1957) and Parnes (1961). The following questions and instructions were read to the group as the discussion proceeded:

1. What exactly is the problem? What has been happening? Why? How serious is this? (To *L*: 5 minutes or less.)

2. What might be done to alleviate this problem? Please mention every idea you can think of that might help to (reduce this loss and destruction of library materials). During this phase of the discussion there must be *no criticism*: all criticism is ruled out. *The wilder the ideas the better*. Even off-beat, impractical suggestions may trigger in other group members practical suggestions which might not otherwise occur to them. *Quantity is wanted*: the greater the number of ideas, the greater the likelihood of good ones. *Combination and improvement are sought*: suggest how ideas given by others could be turned into better ideas, or how two or more ideas can be combined into a better idea. (To *L*: for the next 10 minutes write the gist of all ideas on the blackboard. At any sign of criticism say this: "Please, no judging now.

We'll evaluate these ideas later." Avoid any expression of judgment on your part. If group runs dry, ask: "Are there any more ideas? Any other types of ideas?")

3. Now, how are we going to judge among ideas? What criteria, standards, or characteristics should we consider when choosing the best idea? (To *L*: allow 5 minutes for suggesting and discussing criteria. List these on the blackboard. When no one says anything more, sum up and ask if there are more criteria, unless the 5 minutes are gone.)

4. Now let's evaluate our ideas to pick out the best ones. I'll go down the list. As I do so, tell me whether to rate each one E for excellent, G for good, F for fair, or P for poor. (To *L*: allow up to 10 minutes for this. If group can't agree on an idea, go on to next, then return if time.)

5. What will we report as the best solution for the problem? (To *L*: this may have already been decided, but in no case should you allow more than 5 minutes for picking the final solution.)

Pattern B, *criteria-ideas*, closely approximates the reflective-thinking pattern recommended in the majority of discussion texts. The only difference in B from A is the inversion of Steps 2 and 3.

Pattern C, *problem-solution*, is based on the phase movements discovered by Bales and Strodtbeck (1951) as adapted by Howell and Smith (1956, pp. 51-55):

1. What exactly is the problem? What is happening? Why? How serious is this problem?

2. How should we solve this problem? What are the relative merits of the ideas we can think of? (To *L*: allow up to 25 minutes for suggesting and discussing ideas. If group runs dry before then and has not reached a decision, ask: "Are there any other ideas we ought to discuss?" Record ideas but not evaluations on the blackboard.)

3. What do we report as the best solution for the problem? (To *L*: allow up to 5 minutes additional time for this step if needed.)

Measures

After each discussion Ss completed a six-item questionnaire. Responses to Questions 1 and 2 were on a Likert-type scale, with five steps from "I agree" to "I disagree."

1. Our group arrived at the ideal solution to the problem.

2. Our group arrived at the best solution we could have achieved.

3. Among the ideas that were proposed, I think our group chose: the best one; one of the better ones; an average one; one of the poorer ones; a very poor one.

4. The solution I personally believe to be the best is _____.

5. Check the term which best expresses your *degree of satisfaction* with the *designated leader* of the discussion: very satisfied; somewhat satisfied;

neither particularly satisfied nor dissatisfied; somewhat dissatisfied; very dissatisfied.

6. Check the term which best expresses your *degree of satisfaction* with the *procedure* followed during the discussion: (as for 5).

After the third round, one additional question was asked:

7. Three different patterns for problem solving were used in the three discussions. They were: (a) problem—possible solutions—criteria—evaluation—decision; (b) problem—criteria—possible solutions—evaluation—decision; (c) problem—solution. If you were to lead the discussion of a similar problem, which of the three sequences would you: _____ most prefer to use? _____ least prefer to use?

Responses to Questions 1, 2, 3, 5, and 6 were given numerical values of 5 to 1 for computation. Responses to 7 were ranked 1, 2, 3.

The tentative solutions, final solutions, criteria, and time were obtained from forms filled out by the recorders. The number of good ideas was obtained by a process described by Taylor, Berry, and Block (1958, pp. 40-42). Lists were compiled of all ideas suggested for each problem, and the ideas then grouped into categories. The total numbers of ideas were 270, 300, and 251, respectively. The writers, working independently, rated each idea on five-point scales for (a) feasibility, (b) effectiveness, and (c) generality. All ideas receiving at least two ratings of four and one of three by both raters were called good. Ideas rated good by one rater but not by the other were discussed, and a decision made to classify the ideas as good only if agreement could be reached. Initial agreement was found on 67%, 69%, and 66% of all ideas for the three problems.

RESULTS

Products of Discussions

Final Solutions. The 27 final solutions to each problem were randomly listed. The solutions in each list were then ranked by two experts in the problem area: the experts for the *library problem* were the University Librarian and Assistant Librarian; on the *travel problem* ranks were assigned by an Assistant Professor of Advertising and a Visiting Professor of Advertising and Promotion; on the *teacher problem* ranking was done by two Professors of Secondary Education both of whom have done special study of teacher supply.

Rho correlations between pairs of ranked lists were $+0.36$ ($p < .05$), $+0.34$ ($p < .05$), and $+0.28$ ($p < .10$). Great similarity among solutions may have lowered the correlations.

The ranks were divided into thirds, and

TABLE 1
ANALYSIS OF VARIANCE AMONG NUMBERS OF SOLUTIONS
FOR NIGHTS, PROBLEM-SOLVING PATTERNS, LEADERS,
AND INTERACTIONS

Source	SS	df	σ^2	F	p
Night	43.62	2	21.81	1.41	—
P-S pattern	501.56	2	250.78	16.18	<.01
Leader	276.44	8	34.57	2.23	<.10 ^a
Night X Pattern	78.37	4	19.59	1.26	—
Night X Leader	330.59	16	20.66	1.33	—
Pattern X Leader	187.33	16	11.71	.78	—
Error	496.07	32	15.50 ^b		

^a Almost significant at a p of .05 (2.25).
^b Since there is only one case per cell the error variance is identical with a second-order interaction.

differences among numbers of solutions in each third for each pattern were compared by chi square. The chi square of 3.88 was not significant at the .05 level of confidence.

Numbers of Ideas. The analysis of variance among different problem-solving patterns, among *Ls*, among nights, and for first-order interactions is reported in Table 1. When reading Table 1, it is important to recall that each group encountered the three problems in the same order and that a rotation of problem-solving patterns guaranteed that each problem was discussed nine times using each pattern.

There was a surprising similarity in mean numbers of solutions proposed for the problems: 14.0, 12.3, and 13.6, respectively. The problem-solving patterns incorporating the principle of deferred judgment or brainstorming yielded approximately 50% more ideas than did the pattern in which ideation and evaluation are combined: 15.1 for Pattern A, 15.0 for Pattern B, and 9.8 for Pattern C.

The mean number of ideas per problem per leader was remarkably constant. Except for two leaders, the range was 12.2 to 14.8 (groups led by these two leaders averaged 10.1 and 17.1 ideas per discussion). Differences among leaders for different problems and patterns were not significant at the .05 level of confidence.

Numbers of Good Ideas. The analysis of variance among numbers of good ideas tended to support the previous finding. None of the *F* ratios was significant at the .05 level, but the *F* ratio among problem-solving patterns was almost significant (*p* < .10). The means

were 4.37 for Pattern A, 3.67 for Pattern B, and 3.11 for Pattern C. Means tests revealed that Pattern A yielded significantly more good ideas than did Pattern C (*t* of 2.85, *p* < .01).

Satisfaction with Solutions, Leaders, and Patterns. Answers to Items 1, 2, 3, 5, and 6 on the questionnaires were compared among groups, patterns, and nights. For this analysis the raw data were sums of ratings for all five *Ss* on a given scale. None of the *F* ratios approached significance for either main effects or interactions.

Choice of Problem-Solving Pattern. Following the third round of discussions, *Ss* were asked to indicate which of the three problem-solving discussion patterns they would most and least prefer to use if leading a similar discussion. There was a slight and consistent tendency to both choose and reject (least prefer) the pattern encountered first, but the differences were not even close to significant.

As Table 2 shows, Pattern B was chosen somewhat less often and rejected much more frequently than either Pattern A or C. Differences among numbers of *Ss* choosing and rejecting each of the three patterns on each of the three nights were not significant (chi square of 3.76, *p* < .50 for choosing; chi square of 2.35, *p* < .50 rejecting).

Responses to Question 7 were used to determine rank preference for the three discussion patterns. The differences shown in Table 2 are highly significant (chi square of 25.156, *p* < .001).

An additional test was made of differences among numbers of *Ss* most and least preferring each pattern, omitting Rank 2. The chi square of 10.56 was significant (*p* < .01). Differences in numbers of *Ss* most and least preferring Patterns A and B (chi square 8.14, *p* < .01) and Patterns B and C (7.08, *p* < .01) were also significant. There was no significant difference between numbers of *Ss* most and least preferring Patterns A and C.

DISCUSSION

Much argument has revolved around the procedures called brainstorming in which the processes of ideation and evaluation are separated, but all too little information regarding the effect of such procedures on

TABLE 2

NUMBERS OF SUBJECTS WHO MOST PREFERRED, SECOND MOST PREFERRED, AND LEAST PREFERRED EACH OF THREE PROBLEM-SOLVING PATTERNS

Rank preference	Pattern			Total
	A	B	C	
First	46	40	49	135
Second	57	29	49	135
Third	32	66	37	135
Total	135	135	135	

group problem-solving discussion has been published. The present study showed that with the type of problems and experimental groups used, brainstorming tends to produce both more tentative solutions and more good tentative solutions than does a pattern of discussion in which evaluation is combined with ideation. Evidently the emphasis usually placed on value and quality during a problem-solving discussion can dampen the expression of ideas of potential merit.

That two detailed patterns for group problem-solving discussion produced significantly more ideas than the simpler problem-solution pattern indicates that there may be some advantage in training discussion and conference leaders to use the more detailed patterns, and especially to separate ideation and evaluation. Experimental findings (Maier and Hoffman, 1962) have previously indicated that ideas produced later in the problem-solving process tend to be superior to those produced first. Methods for delaying decision-making appear to have merit for both individuals and groups.

That significantly more Ss expressed a greater preference for an ideation-criteria-evaluation pattern or a simple problem-solution pattern than for a criteria-ideation-evaluation pattern indicates that talking about criteria apart from definite ideas to be judged may be artificial and frustrating. The advice given in the majority of current discussion textbooks and manuals to establish criteria before attempting to find solutions appears dubious at best and harmful at worst. Teaching a pattern based on speculation and casual observation apart

from experimental investigation may have lowered both productivity and satisfaction in many conferences and discussions.

The lack of significant differences on ratings of leadership, procedures, and solutions may indicate either that the scales were too crude to differentiate or that other factors had more influence on satisfaction and commitment than did the patterns. Possibly the effects of group power, leadership style, and similarity in the quality of final solutions masked the impact, if any, of problem-solving patterns on satisfaction.

The findings reported herein may apply only to the type of problem with which discussants are not directly involved. The findings do not necessarily apply to discussions by persons who will be directly and immediately affected by the products of their deliberation. Nor was any attempt made to consider the effect of failing to explore a problem before attempting to solve it. These matters beg further investigation.

REFERENCES

- BALES, R. F., & STRODTBECK, F. L. Phases in group problem-solving. *J. abnorm. soc. Psychol.*, 1951, 46, 485-495.
- BARNLUND, D. C., & HAIMAN, F. S. *The dynamics of discussion*. Boston: Houghton Mifflin, 1960.
- DEWEY, J. *How we think*. New York: Heath, 1910.
- GOLEMBIEWSKI, R. T. *The small group*. Chicago: Univer. Chicago Press, 1962.
- HOWELL, W. S., & SMITH, D. K. *Discussion*. New York: Macmillan, 1956.
- MAIER, N. R. F., & HOFFMAN, L. R. Group decision in England and the United States. *Personnel Psychol.*, 1962, 15, 75-87.
- MAIER, N. R. F., & MAIER, R. A. An experimental test of the effects of "developmental" vs. "free" discussions on the quality of group decisions. *J. appl. Psychol.*, 1957, 41, 320-323.
- OSBORN, A. F. *Applied imagination*. New York: Scribners, 1957.
- OSBORN, A. F. Developments in creative education. In S. J. Parnes & H. F. Harding (Eds.), *A source-book for creative thinking*. New York: Scribners, 1962.
- PARNES, S. J. *Student workbook for creative problem-solving courses and institutes*. Buffalo: Univer. Buffalo Bookstore, 1961.
- TAYLOR, D. W., BERRY, P. C., & BLOCK, C. H. Does group participation when using brainstorming facilitate or inhibit creative thinking? *Admin. Sci. Quart.*, 1958, 3, 23-47.

(Received May 9, 1963)

VOLUNTARY FEMALE CLERICAL TURNOVER: THE CONCURRENT AND PREDICTIVE VALIDITY OF A WEIGHTED APPLICATION BLANK

WILLIAM D. BUEL

Vernon Psychological Laboratory, Chicago

Through differential weighting of application-blank information ($N = 152$), 2 validations against voluntary female clerical-turnover criteria were accomplished. Concurrent (16 scoreable items with $N = 72$) and predictive (13 scoreable items with $N = 60$) cross validities are reported on holdout and follow-up samples, respectively. Statistically significant relationships are demonstrated in both the concurrent and the predictive studies despite intervening variables, preceding the predictive study, sufficient to destroy validity.

Psychologists have frequently demonstrated the validity of application blank and biographical information for assessing predispositions such as potential for salary increase (Scollay, 1956), tenure or turnover (Minor, 1958), credit risk (McGrath, 1960), and research competence and creativity (Smith et al., 1961). The large majority of the published testimonials to this validity have, however, centered in demonstrations of concurrent, and somewhat less frequently predictive validity, but rarely both. The present study is perhaps unique in that it reports both concurrent and predictive validity for the same weighted application blank in two different environs for two different samples. In so doing, it reflects the degree to which a weighted application blank retains validity over time in spite of intervening variables which generally tend to limit or destroy validity.

METHOD

In order to develop a weighted application blank which would be predictive of voluntary female clerical turnover, a personal action (voluntary termination) criterion of turnover was necessary. To establish such a criterion, personnel records were searched for the application blanks of individuals in the job classes, stenographer, typist, and clerk who had joined the general office of a major petroleum company at least 3 years prior to the date of the search and who were still employed.¹ Records were also searched for the application blanks of individuals who had joined the organization sometime between

January 1, 1956 and the date of the search but who had voluntarily left the organization prior to the completion of 9 months service.² Individuals who left via discharge or for reasons of pregnancy were deleted from the sample.

From these two samples, 224 blanks were selected which conformed to one of the criteria above. This group was further divided into one subgroup of 152 blanks (73 short tenure and 79 long tenure) for item-weighting purposes and another subgroup of 72 blanks (a holdout group of 36 short tenure and 36 long tenure) for purposes of concurrent cross-validation.

By virtue of being reasonable of quantification or categorization, 35 items of information were selected for statistical analysis. Employing the weighting sample, simple frequencies of category response were tallied for both the long- and short-tenure groups, and these frequencies and their intracategory proportional values examined for differences. On an "appearance" basis, 16 of the items were chosen for differential weighting by the method suggested by Wherry (1944), that is, no statistical tests were employed to identify significant differences between intracategory response proportions. These 16 items and their weights were amalgamated into a single-overlay scoring key.

The weighted items represented typical bits of application-blank information—distance from home to office, work best qualified for, acquaintances in the company, who referred applicant to the company, marital status, length of time married, schooling, participation in school-sponsored sports, participation in class organizations, participation in other school-related activities, how the applicant spent summer vacations, references, type of residence, how long applicant lived in city, and age. No effort was made, as suggested by Mahoney (1958), to limit the number of or accord separate treatment to contingency items. In the truest sense, only one item (length of time married) is contingent. While educational items are frequently contingent, those included here, being in-

¹ The author wishes to thank R. L. Larson and the Pure Oil Company for permission to report these studies.

² Both searches took place in January of 1959.

dependent of level of schooling or graduation, are not. However, since both validities to be reported here are cross-validated, the spurious effect of contingency (weighting multiple pieces of correlated information), if it existed, would be minimized if not completely eliminated.

RESULTS

Concurrent Cross-Validity

To determine the validity of the weighted information in total-score form, the 72 hold-out blanks were analyzed with the 16-item key and the resulting total scores correlated with the turnover criterion. A biserial correlation from widespread classes (Peters and Van Voorhis, 1940) of .49, significant at the .01 level of confidence, was obtained.³ For purposes of future application by the employment department, a cutting score of 10 (the point of maximum separation between long- and short-tenure cross-validation groups) was assigned.

Intervening Variables

Had it not been for the following restrictions, the screening method described would have been put into effect in the late spring of 1959. However:

1. During the course of the study a decision was made to move the general office of the organization approximately 25 miles into the suburbs. This threatened to lessen if not destroy the validity of the key.

2. A tight labor market was anticipated, and subsequently experienced, as a function of the organization's not being attractive to potential clerical employees who would not or could not move to a suburban location.

3. It was feared that, once recruiting began in the suburbs in anticipation of the move,

³ The statistical treatment employed corrects for the fact that the criterion groups were separated by a void between 9 months and 3 years service and would have, if not specially treated, given rise to a spuriously high biserial correlation.

Age, as a function of the differing time periods from which the criterion groups were drawn, was impossible of control in the weighting sample. In the concurrent cross-validation sample, however, the mean ages of the long- and short-tenure groups were 26.9 and 28.0 years, respectively, controlling the possibility of an age-stability interaction benefit for the long-tenure group.

a different and invalidating sample of job applicants would be encountered.

Fearing inapplicability, the weighted application blank was sidelined as a screening device for the period of June, 1959 to December, 1960. As might be imagined, turnover for reasons atypical of the weighting sample was great during this period because of the necessary transition from an urban to a suburban clerical force.

The geographic move took place in October of 1960. Because of the massive restaffing job that occurred in the period from approximately 3 months before to 1 month after the move, a large number of new blanks were available and possible of comparison with the population upon whom weights were originally developed. The comparison suggested that, with regard to the weighted items, there was not as large a population difference as had been anticipated, partially dispelling the previous concern. It was therefore decided to reinstitute the weighted application blank as a screening device, while reducing the cutting score from 10 to 8 because three items, with a mean combined contribution to total score of 1.84, were unreasonable of scoring in the face of a geographic move. The weights for the 13 remaining items were those derived in the original weighting effort.

Predictive Cross-Validity

The application blanks for the follow-up study were drawn from the file of hires for the period of January 2, 1961 through December 1, 1961, although determination of which group a blank belonged to was not made until December 1, 1962. The reason for not utilizing the blanks of persons who joined the organization after December 1, 1961 was that even those who were to remain employed indefinitely would have been on the job less than 1 year at the time of the study, and it was deemed unwise to wait any longer before checking the efficiency of the procedure. Short tenure was therefore defined as 1 year or less and long tenure as 1 year or more with a maximum of 23 months possible for those hired at the begin-

ning of the study period.⁴ Short tenure was not defined as 9 months or less because N would have been reduced by disallowing 3 months of criterion time.

Sixty persons were identified in the time period described. Forty were members of the long-tenure group and 20 were members of the short-tenure group. The biserial correlation between weighted application-blank scores and the tenure dichotomy was .33, significant at the .02 level of confidence.

DISCUSSION

The concurrent validity portion of this study, because of the relatively typical procedures and methods of reporting results, needs little discussion. Reconsideration of the restrictions operating against predictive validity, however, shows that: three items of the original 16 were lost because of the move; the cutting score was reduced (selection ratio eased since the cutting score was reduced by more than the mean contribution of the three deleted items); a new or at least potentially different population was encountered in the predictive situation; the instrument was used in selection and thereby a restriction in range in the form of pre-selection was imposed upon the predictor scores; and the criterion was redefined in such a way as to restrict variance. Each of these restrictions alone had the potential for lessening or destroying validity. That they did so only partially (the validity shrinkage demonstrated is perhaps of the approximate magnitude to be expected in predictive cross-validation situations free of such restrictions) is encouraging and may lend support to the oft enunciated stability and reliability advantages of biographical data.

Some readers may question why a valid screening device allowed 20 out of 60 persons to be selected into the low criterion group, a figure which is actually higher than the industry average for female clerical turnover (20 to 25%). The 33% reported here is more apparent than real. As most personnel

practitioners are aware, and as demonstrated by the fact that the mean length of service within the short-tenure portion of the predictive cross-validation group was 5.1 months, much clerical turnover takes place shortly after employment (within 3 to 6 months). Turnover is typically reported by the year but reflects persons who have been employed for multiple years and, through time, have the effect of minimizing the apparent yearly turnover percentages and maximizing mean length of tenure. By their very definition and nature, such yearly reports do not include previous years' losses (short tenure), but do include previous years' acquisitions and retentions (long tenure). Fixed to a hiring date, these data, encompassing a relatively short span of time, took the brunt of immediate turnover and therefore, quite spuriously, suggest above average turnover.

A re-analysis of the entire application blank and a re-weighting of these and possible other items has been recommended, based upon the assumption that improved validity could be achieved by weights more specific to the suburban population. At such time as further study might be accomplished, a rescoring of both the holdout and follow-up samples and a comparison with new holdout and follow-up samples should be possible. Such comparisons should lend insight into reliability and validity as a function of type of information weighted and of population characteristics.

REFERENCES

- MCGRATH, J. J. Improving credit evaluation with a weighted application blank. *J. appl. Psychol.*, 1960, **44**, 325-328.
- MAHONEY, T. A. Weighted application blank analysis of "contingency" items. *J. appl. Psychol.*, 1958, **42**, 60-62.
- MINOR, F. J. The prediction of turnover of clerical employees. *Personnel Psychol.*, 1958, **11**, 393-402.
- PETERS, C. C., & VAN VOORHIS, W. R. *Statistical procedures and their mathematical bases*. New York: McGraw-Hill, 1940.
- SCOLLAY, R. W. Validation of personal history items against a salary increase criterion. *Personnel Psychol.*, 1956, **9**, 325-335.
- SMITH, W. J., ALBRIGHT, L. E., GLENNON, J. R., & OWENS, W. A. The prediction of research competence and creativity from personal history. *J. appl. Psychol.*, 1961, **45**, 59-62.
- WHERRY, R. J. Maximal weighting of qualitative data. *Psychometrika*, 1944, **9**, 263-266.

(Received May 27, 1963)

⁴ Since residence in the long- or short-tenure groups was defined, in the follow-up study, on an either-or basis (in the long-tenure group 21 months employment was treated no differently than 13 months), no restriction is introduced as a function of an individual being in the long-tenure group but not having joined the organization soon enough to have a maximum criterion score.

REMOTE ASSOCIATES TEST AS A PREDICTOR OF CREATIVITY IN ENGINEERS

LOIS-ELLIN DATTA¹

General Electric Company

A correlation of +.31 was found between Remote Associates Test scores and supervisory ratings of creativity for 21 American-born engineers. Production of remote verbal associates may not be as important a component of behavioral creativity for professional engineers as it may be for other groups.

The capacity to develop unusual ideas which also meet some criterion of adaptiveness is central to many definitions of creativity. Mednick measures one aspect of this capacity by the Remote Associates Test (RAT). In this test, the subject (*S*) is asked to find a fourth word which is related to three other words. The three other words are apparently unrelated among themselves; e.g., the remote associate of "balloon," "soda," and "Dad" is "pop."

Mednick (1962) describes a number of studies indicating the relationship of creativity and uniqueness of associations, including a correlation of +.70 between faculty ratings and RAT scores for 21 design students. The present study reports a second attempt at industrial validation of the RAT. In a previous study with 31 physicists, the nonsignificant correlation (+.13) between RAT and creativity was attributed to language difficulties for the more creative scientists. The data from this second attempt indicate little improvement in predictive validity when all *Ss* speak American-English as their native language.

Each of 21 engineers working in the Quality Control and Measurements Operation was rated jointly by his immediate supervisor and by the section manager. The ratings were made on a modification of Taylor's (1962) creativity scale for National Aeronautics and Space Administration scientists. Although reasonably good distributions were

obtained for both the creativity ratings and RAT scores, the rho between the two measures was only +.31, which is not significant at the .05 level of confidence with $df = 19$.

There was no evidence that cultural or linguistic variables were suppressing a higher "true" relationship. All *Ss* were native Americans and linguistic fluency, as measured by the number of words beginning with *S* and ending with *T* which were written in 1 minute, did not differentiate high and low RAT scorers. A third possibility is that the highest rated engineers were not sufficiently creative for the RAT to discriminate or that the highest RAT scorers were more typical of less creative *Ss* in other groups. Mednick has not made RAT data available for comparison, but it should be noted that a correlation he cites between rated originality and RAT scores for 40 eminent architects is similar in magnitude (+.31) to the relationship found in the present study.

The data suggest that the production of remote verbal associations is not as important a component of behavioral creativity for professional engineers (and perhaps architects and scientists) as it may be for psychology and design (and perhaps other) students.

REFERENCES

- MEDNICK, S. A. The associative basis of the creative process. *Psychol. Rev.*, 1962, **69**, 220-232.
TAYLOR, C. W. Rating scales for NASA scientists. Paper read at the University of Utah research conference on the identification of creative scientific talent, Brighton, Utah, 1962. (Suppl. material)

(Received May 27, 1963)

¹ Work done while at General Electric Company. Now with the Laboratory of Psychology, National Institute of Mental Health, Bethesda, Maryland.

A NOTE ON THE REMOTE ASSOCIATES TEST, UNITED STATES CULTURE, AND CREATIVITY

LOIS-ELLIN DATTA¹

General Electric Company

An industrial validation of a measure of creative potential, the Remote Associates Test (RAT), found the correlation between RAT scores and supervisory ratings of creativity for 31 physicists to be $+ .13$, with 6 of the 10 high-rated scientists having very low RAT scores. These scientists do not speak English as their native language although their linguistic fluency was no different from that of the other scientists. The proportion of nonEnglish-speaking scientists in this sample is similar to that reported for eminent mathematicians; the RAT may be limited as a measure of creative potential in this occupational group.

This note reports an attempt at industrial validation for a new measure of creative potential, the Remote Associates Test (RAT).

Central to many definitions of creativity is the achievement of original and worthwhile ideas. By defining original as statistically infrequent and worthwhile as meeting certain parameters of meaning, Mednick (1962) has developed a measure of the ability to demonstrate unusual and satisfactory associations. In each item of this test, three apparently unrelated words are given and the subject (*S*) is required to name a fourth word which "is related to all three"; for example, "balloon," "soda," and "Dad" have "pop" as an associate.

Mednick reports that performance on the RAT predicts the judged creativity of psychology and design students, and of architects; research is currently under way with a number of other groups. This note reports, however, an instance in which RAT scores did not predict rated creativity; further inquiry into the results suggests a limitation of the RAT for certain occupational groups.

The *Ss* were 31 research scientists (physicists, physical chemists, mathematical physicists) in a large industrial space sciences laboratory. Each scientist was rated by his immediate supervisor on a scale of creativity derived from Taylor's measure (1962). The rho between RAT scores and rated creativity

was $+ .13$. An examination of the distribution indicated a peculiar split in the RAT scores of the 10 scientists in the upper third of the rated creativity distribution. Four of these men achieved scores in the upper or middle thirds of the RAT distribution for the sample, but the scores of the remaining six scientists were extremely low. During interviews with *Ss*, it became apparent that these six scientists represented an instance in which the individual association pairs, (e.g., soda—pop; balloon—pop; Dad—pop) were not as familiar to all *Ss* in this study as they were to "most individuals in the culture."

A third of 31 scientists in the total group do not speak English as their native language (including two *Ss* from close-knit subgroups within the United States culture). Six of these 10 scientists were rated as highly creative; these were the same six scientists in the high creative group whose RAT scores were extremely low. Although the verbal fluency scores of the 10 "non-United States cultured" *Ss* were not significantly lower than those of their colleagues, their median RAT score was 11.0, compared to the 20.0 of the other 21 scientists.

These results suggest that the RAT may be limited as a measure of one aspect of creative potential in physicist mathematician personnel, particularly since this occupational group has been found to contain a higher proportion of creative, foreign-born individuals than the class of scientists in general. Helson (1961), for example, reports that

¹ Work done while at General Electric Company. Now with the Laboratory of Psychology, National Institute of Mental Health, Bethesda, Maryland.

50% of the creative women mathematicians she studied were foreign born and cites Visser's finding that while only 17% of the scientists starred in *American Men of Science* are foreign born, 32% of the eminent mathematicians are not natives of this country.

REFERENCES

- HELSON, RAVENNA M. Creativity, sex and mathematics. In, *The creative person*. Berkeley, Calif.: Institute of Personality Assessment and Research, 1961.
- MEDNICK, S. A. The associative basis of the creative process. *Psychol. Rev.*, 1962, **69**, 220-232.
- TAYLOR, C. W. Rating scales for NASA scientists. Paper read at the University of Utah research conference on the identification of creative scientific talent, Brighton, Utah, 1962. (Suppl. material)
- VISHER, S. S. *Scientists starred 1903-1943 in "American Men of Science": A study of collegiate and doctoral training, birthplace, distribution, backgrounds, and influences*. Baltimore, Md.: Johns Hopkins Univer. Press, 1947.

(Received May 27, 1963)

THE COMPREHENSIBILITY OF SEVERAL GRAMMATICAL TRANSFORMATIONS¹

E. B. COLEMAN

Sul Ross State College

4 experiments compared the comprehensibility of different grammatical transformations of a passage. In 2 experiments, difficult prose was simplified by transforming nominalizations, adjectivalizations, and passives to their active-verb transforms. In the other 2, nominalizations alone were compared to their active-verb transforms. In all 4 experiments—which used several different presentation modes and several different dependent variables—the active-verb transforms were found to be more comprehensible.

The notion of altering sentences so as to make them more readable implies that it is possible to alter form with a minimum of change in meaning, but sentences that differ even slightly in form probably also differ slightly in meaning, i.e., they represent slightly different ideas. Therefore the experimenter who is comparing the readability of different forms of a sentence should describe the forms precisely enough that a writer who wishes to use the more readable one can decide whether it will represent his idea with acceptable precision.

Linguists (e.g., Chomsky, 1956; Harris, 1957; Lees, 1960) have provided formulas called grammatical transformations that describe certain alterations in form rather precisely. For instance, passives (*It was described by John*) and nominalizations (*John's description of it*) are grammatical transformations of the active-verb form (*John described it*). An inspection of a list of transformations of a single sentence gives a strong impression that some are more readable than others. The basic purpose of the present paper is to introduce the use of transformations as independent variables in studies of readability.

The experiments reported in the present paper are exploratory and nonanalytical; even so, they have immediate and obvious implications for anyone interested in improving readability. In the first two experiments, three transformations were applied simultaneously to difficult technical prose (nominalizations, passives, and adjectivalizations were trans-

formed to their active-verb counterparts), and the original versions were compared to the transformed versions for readability. The last two studies were slightly more analytical; nominalizations alone were compared to their active-verb transformations.

EXPERIMENT I

Procedure

Experimental Design. Two long prose passages were simplified by applying three transformations to them, and the original versions were compared to the simplified versions using multiple-choice tests and a Lindquist Type V design (Lindquist, 1953, p. 288). This design consists of several Graeco-Latin squares so that it allowed order of presentation, subjects (Ss) and passages to be counterbalanced. The most essential point is that both passages were typed in both versions (original and simplified), and each S read one passage in the original and the other in its simplified transformation.

Subjects. The Ss were 48 students enrolled in an introductory psychology course at Johns Hopkins University.

Materials. Two difficult passages, each 2,969 words in length, were selected from two articles reprinted in Cartwright and Zander (1953): (a) "A Theoretical Framework for Interaction Process Analysis" by R. F. Bales; and (b) "The Dynamics of Power" by R. Lippit, N. Polansky, F. Redl, and S. Posen. The passages were closely matched in difficulty as measured by the Flesch Reading Ease Score (1949), that of the first being 40, and that of the second being 43.

Both passages were simplified by applying three grammatical transformations: (a) Passive verbs were transformed to actives; e.g., *The car was bought* → *He bought the car*. See Harris (1957, p. 325) for a more complete description of this transformation and its formulas. (b) Nominalizations using abstract nouns were transformed to their active-verb derivatives; e.g., *His explanation of the design . . .* → *He explained the design. . .* (c) Adjectivalizations were transformed to their adjectival or adverbial forms; e.g., *He has social power* → *He is socially powerful*.

¹ This research was supported by a grant from the National Science Foundation, GB-241.

See Lees (1960, especially Chapter 3) for a more complete description of transformations involving nominalizations and adjectivalizations.

The following details are worth listing: (a) The simplified versions had almost exactly the same content morphemes² as the nominalized versions, i.e., the vocabulary was not watered down to a less technical one. The major exception was that one transformation necessitated discarding circumlocutions such as: *engaged in interaction* → *interacts*; *coming to an agreement* → *agrees*; etc. In addition, seven sentences necessitated such minor variations as the following: *the first study was concerned with* → *the first time we studied*; *Our initial curiosity focused upon* → *Initially we were curious about*; *We measured the degree of success* → *We measured how often he succeeded*. (b) 4.2% of the nominalizations could not be conveniently transformed. Another 9.8% had to be transformed to gerunds or infinitives rather than to full verbs; (c) In general, such simplified transformations have fewer words. To reduce this difference, articles and prepositions were used lavishly in the simplified versions, sometimes to the point of awkwardness, but they still contained 1.6% fewer words than the nominalized versions.

Presentation. After reading the instructions, *S* read a 232-word warm-up selection and took a four-item multiple-choice test on it. Then he read the two passages typed on ordinary typing paper—one in the original and the other in the simplified version. He was given 12 minutes to read each passage; and immediately after finishing a passage, he took its test. He was scored on number of words read and number of questions answered correctly. He did not answer questions on material he had not reached in his 12 minutes.

Multiple-choice Questions. It is unlikely that the questions could be biased for either version because both versions used almost exactly the same content morphemes. But to guard against any possible bias, the following procedure was used: I composed about 30 questions for each passage in rough outline form. A graduate student was given one passage in the original version and the other in the rewritten version, and asked to reword all questions. "Rewording" might be complete reorganization or only a change in verb tense. A second graduate student did the same with the other versions of the passages. These two sets of rewritten questions were then given to a colleague who had never seen the passages, and he (again rewording) composed the final set of questions—18 multiple-choice questions and one matching question for each passage.

² As a working approximation, content morphemes can be considered to be the words that are capitalized in titles minus their inflectional and derivational suffixes. Function morphemes can be considered to be the uncapitalized words plus inflectional and derivational suffixes. For a more extensive discussion, see Hockett's *contentives* and *functors* (1958, p. 264).

Results

Thirty *Ss* answered more questions about the simplified versions, 11 answered more about the originals, and there were seven ties. By a binomial test, a ratio of 30 to 11 is significant beyond the .005 level. There was no significant difference in mean number of words read, 2,169 for the simplified versions and 2,160 for the originals.

Anyone interested in improving readability would be heartened by the magnitude of the improvement. When the mean questions answered correctly are corrected for guessing, the means are 5.38 for the simplified versions and 4.29 for the originals, thus the magnitude of the improvement is 25.2%. This estimate of magnitude is not very reliable because it will vary according to the relation between the intelligence of the readers and the difficulty of the passage. Still this improvement is heartening because the only changes made were in the grammatical frame of function morphemes: The content morphemes were not diluted to less technical synonyms.

EXPERIMENT II

Procedure

Experimental Design. Experiment II was simply a replication of Experiment I that used a wider sample of passages. The design was identical to Experiment I except that each *S* read four passages in the original version and four in the simplified version.

Subjects. The *Ss* were 16 undergraduates from Johns Hopkins University.

Materials. Eight different paragraphs that ranged in length from 86 to 117 words were selected from undergraduate texts. They were simplified by applying the same three transformations used in Experiment I, and typed on ordinary typing paper.

Presentation. The *Ss* were given .5 second per word to read the passages. It should be noted that all eight of the transformed versions were shorter than the originals, and thus insofar as the transformations held content morphemes constant, *S* had less time to learn these morphemes in the simplified versions.

Scoring. Immediately after reading a paragraph, *S* was instructed to write it to the best of his ability. He was instructed to write it exactly if possible, but to use synonyms if he could not recall the exact words, and to write every individual word he could remember even if he could not remember how it was used in the paragraph. Recall was scored in four ways: (a) number of content words correctly reproduced; (b) number of content words correctly reproduced plus number of synonyms for content words; (c) number of content words in correct kernel sentences (that is, if *S* wrote a sentence that

TABLE 1
MEAN CORRECT SCORES FOR RECALL OF NOMINALIZED
PARAGRAPHS VERSUS RECALL FOR SIMPLIFIED
VERSIONS—WITH WILCOXON *T*s

Criterion	Original, nom- inal- ized	Simpli- fied	Wilcoxon <i>T</i>
Content word	43.4	45.8	4****
Content word plus synonym	61.0	65.6	48
Content word in correct kernel	33.4	37.4	11****
Content word plus synonym in correct kernel	48.7	52.4	27**

** *p* ≤ .025, one-tailed.
**** *p* ≤ .005, one-tailed.

was wrong or irrelevant, it was reduced to kernel sentences and all kernels that did not match the kernels of the presented paragraph were discounted in totaling his score); ³ (*d*) number of content words plus synonyms for other content words in correct kernels. In all four of the above scores, any derived or inflected form of a word was counted correct.

Results

The simplified versions were recalled more accurately than the originals for all scoring systems, although one system did fail to reach an acceptable level of significance (see Table 1).

EXPERIMENT III

Procedure

Experimental Design. Experiment III (and IV) compared nominalizations to their transformations using active verbs. A Lindquist Type V design was used to compare 20 nominalized sentences with their active-verb transformations.

Subjects. The *Ss* were 20 undergraduates from Johns Hopkins.

Materials. Twenty sentences that contained an abstract noun nominalized from a verb were selected by random sampling and simplified by transforming

³ The reader may gain some understanding of both kernels and the extent to which these transformations hold meaning constant if he notes that both versions reduce to the same kernels. For example, *John's operation of the large boat was skillful* or *John operated the large boat skillfully* both reduce to the same kernels: *John operated the boat, This was skillful, The boat is large.* For a further discussion of kernels, see Chomsky (1957, p. 45) and Harris (1957, p. 335).

the nominalization to an active verb. It should be made explicit that the set of sentences as well as the *Ss* is a sampling variable. That is, the null hypothesis can be rejected for the sentence population as well as *S* population. The sentences were randomly selected from the library at Johns Hopkins as follows: A drawer was selected from the card file by random numbers and a card was drawn. If the book was not in the stacks, the book to its left was chosen. The selected book was opened to page 100 (page 25 if it had less than 100 pages), and the first sentence containing an abstract noun nominalized from a verb was chosen. Then a second drawer was selected, and so on until all 20 sentences had been selected. There was one restriction. When the original sentence was transformed, the simplified version had to have the same number of words as the original. This necessitated drawing 23 sentences, three being dropped because there was no way to equalize length in the two versions. Sentence modifiers such as *of course* were added to the originals before transforming so that all 40 sentences were 20 words in length.

Presentation. The sentences were typed on 8 by 8 inch flash cards—one sentence to a card. They were exposed to *S* manually for 4 seconds timed by a stop watch. Immediately after a card was withdrawn, *S* was asked to write down all of it he remembered. After finishing his 20 sentences (10 originals and 10 simplified ones) he was given 20 multiple-choice questions—one on each sentence.

Scoring. Performance was scored in four ways: (*a*) number of words correctly reproduced; (*b*) number of content words correctly reproduced; (*c*) number of content words correctly reproduced in correct kernel sentences; (*d*) number of answers correct on the multiple-choice test.

Results

The active-verb transforms were remembered more accurately by every scoring system (see Table 2). All tests of significance were by Wilcoxon matched-pairs tests. The multiple-choice test gave rather disappointing results, failing to reach significance for both samples; however, the difference was in the predicted direction. By all other scoring systems, the differences were significant for both samples—sentences and subjects.

EXPERIMENT IV

Procedure

Experimental Design. Experiment IV was a comparison of 10 nominalized sentences versus their active-verb transformations, and it also used a Lindquist Type V design, except that because many *Ss* were unable to memorize all 10 sentences in the 50-minute testing period, each sentence was memorized by a mean of only six *Ss*.

Subjects. The *Ss* were 18 undergraduates from Sul Ross State College.

TABLE 2

MEAN CORRECT SCORES FOR RECALL OF NOMINALIZED SENTENCES VERSUS RECALL OF SIMPLIFIED VERSIONS—
WITH WILCOXON *T*'S FOR SUBJECT SAMPLE AND
SENTENCE SAMPLE

Criterion	Original, nominal- ized	Simplified	Wilcoxon <i>T</i> , subject	Wilcoxon <i>T</i> , sentence
Word	12.1	14.3	66*	54*
Content word	6.3	7.8	52***	40***
Content word in correct kernel	4.9	6.6	48****	33***
Number of answers	16.1	16.7	104	84

* $p \leq .05$, one-tailed.
 *** $p \leq .01$, one-tailed.
 **** $p \leq .005$, one-tailed.

Materials. Ten sentences were randomly selected as described in Experiment III except that they were chosen so that each contained two nominalizations, both of which were transformed to active verbs in the simplified version. They were typed on tapes for memory drum presentation using a typewriter that types 16 characters per inch.

Presentation. The sentences were presented on a Gerbrand's memory drum at a 1 second rate. A mean of 4.7 words per second was presented to *S*.

Measure. The measure was trials to perfect recall. To increase the number of sentences that *S* could memorize in the testing period, he was given a packet of cards containing all the content morphemes in the sentence and he was asked to arrange these in the correct order and to fill in the framework of function morphemes. For instance, immediately after the first exposure of *The association of written signs with visual images and with auditory signs is only an extension of the same process*, *S* was given a packet of cards on which were typed *associat-, writ-, sign-, vis-, imag-, audit-, sign-, only, exten-, same, proce-*. He was asked to arrange these in the correct order, and if he failed the sentence was presented on the memory drum again, and he tried to improve his ordering. His cards were not scrambled after each presentation. He was informed as soon as his ordering was correct and he then tried orally to fill in the frame of function morphemes. He had available a sheet that contained all function morphemes needed for all sentences, and function morphemes had been defined for him as words that are not capitalized in titles plus word endings. Lack of familiarity with the definition of function morpheme does not seem to have been a handicap because no *S* ever referred to his list of them. It should be emphasized that the transformed and simplified versions of each sentence used the same packet of content morphemes.

Introducing the reconstruction factor may seem an unwarranted complication, so perhaps it is worthwhile to state explicitly the motives for introducing it: (a) It increased the number of sentences that *S* was able to memorize in the experimental period. (b)

It reduced random variance caused by differences in *Ss'* familiarity with the content morphemes. That is, the experimental variable was a change in grammatical frameworks (including *order* of content morphemes) and there are obvious advantages to be gained by insuring that insofar as possible the measured response is a function of the experimental variable alone.

Results

The original nominalized sentences were learned in a mean of 7.61 exposures per sentence, and their active-verb transformations in 6.19. By a binomial test, this difference is significant beyond .015 for the sample of *Ss* (four learned their original nominalized sentences in fewer exposures, and 14 learned the simplified sentences in fewer). Insofar as the 10 nominalized sentences represent a population of such sentences, the above difference is significant for this sample beyond .01 by a Wilcoxon matched-pairs test (*T* was three).

DISCUSSION

The basic purpose of the present paper was to introduce the use of grammatical transformations as independent variables in readability experiments, and all four experiments supported the notion that some transformations are easier to comprehend than others. The last three experiments more specifically suggested that transformations using active verbs are easier to comprehend than their nominalized counterparts.

The linguistic literature provides several explanations as to why nominalizations are relatively hard to comprehend.

1. Jespersen (1924, p. 133–144) noted that nominalized sentences lack many specific references contained in their counterparts using active verbs. Consider, for example, the following nominalization and its counterpart: "AN INCLUSION OF THIS IS AN ADMISSION THAT IT WAS important" → "SINCE SHE INCLUDED THIS, SHE IS admitting THAT IT WAS important." Note that the transformation using the active verbs contains the following specific references that are lacking in the nominalization: subject, person, and number (*she*), past tense (*-ed*), causation (*since*), progressive aspect (*is ———ing*). Of course, these references would be implied in the context sur-

rounding the nominalized sentence, but the transformation using the active verbs expresses them explicitly. To the extent that knowing these references is important in understanding the sentence, and to the extent that the reader finds it difficult to deduce them from the context, he will find the nominalization hard to understand.

2. Flesch has argued that short sentences are relatively easy to comprehend, but a careful reading of his works (1946, p. 32; 1949, p. 129) suggests that he is concerned with clause length more than sentence length. (An experiment by Coleman [1962] also supports the notion that shortening clauses would improve comprehensibility more effectively than shortening sentences.) Transforming nominalizations to active verbs is an effective way to shorten clauses; e.g., a 1,000-word sample from the original passages of Experiment I contained clauses with a mean length of 15.3 words, and their transformed counterparts contained only 8.9 words. It is almost certain that sentence (or clause) length can predict readability only because it is correlated with more fundamental predictors of syntactic complexity such as nesting, transformational complexity, and others (Miller & Chomsky, 1963).

The transformations used in the present experiments varied most of the above predictors simultaneously so only the most general advice for improving readability can be deduced from them. Needless to say, however, the re-

search technique could be refined to perform more analytic studies of syntactic complexity. For instance, Jespersen's notion could be studied by selecting nominalizations that have as many or almost as many specific references as their transformation using the active verb.

REFERENCES

- CARTWRIGHT, D., & ZANDER, A. *Group dynamics*. Evanston, Ill.: Row Peterson, 1953.
- CHOMSKY, N. The logical structure of linguistic theory. Cambridge: Massachusetts Institute of Technology, 1956. (microfilm)
- CHOMSKY, N. *Syntactic structures*. The Hague: Mouton, 1957.
- COLEMAN, E. B. Improving comprehensibility by shortening sentences. *J. appl. Psychol.*, 1962, **46**, 131-134.
- FLESCHE, R. F. *The art of plain talk*. New York: Harper, 1946.
- FLESCHE, R. F. *The art of readable writing*. New York: Harper, 1949.
- HARRIS, Z. S. Co-occurrence and transformation in linguistic structure. *Language*, 1957, **33**, 283-340.
- HOCKETT, C. F. *A course in modern linguistics*. New York: Macmillan, 1958.
- JESPERSEN, O. *The philosophy of grammar*. New York: Holt, 1924.
- LEES, R. B. The grammar of English nominalizations. *Int. J. Amer. Linguist.*, 1960, **26**, No. 3 (Whole Part II).
- LINDQUIST, H. F. *Design and analysis of experiments in psychology and education*. Boston: Houghton Mifflin, 1953.
- MILLER, G. A., & CHOMSKY, N. Finitary models of language users. In R. D. Luce, R. R. Bush, & E. Galanter (Eds.), *Handbook of mathematical psychology*. New York: Wiley, 1963.

(Received June 17, 1963)

A CROSS-CULTURAL STUDY OF ACHIEVEMENT MOTIVATION¹

HARRISON G. GOUGH

University of California, Berkeley

The California Psychological Inventory (CPI) has shown promise in American studies of scholastic achievement, contrasting with typical findings in predicting achievement from personality appraisals. A theoretical issue concerns choice of concepts, whether these should reflect states of distress and disturbance or differential components of interpersonal adequacy. In the latter vein, the CPI emphasizes positive variables of presumably universal relevance. A study of academic achievement in Italy was undertaken, testing 204 males and 137 females from 4 high schools in 3 cities. The achievement motive scales of the CPI (Ac and Ai) correlated $+.32$ and $+.35$ with grades for males and $+.33$ and $+.29$ for females. A multiple regression equation including Ac, Ai, and Fx gave predictive validities of $+.43$ and $+.45$.

Psychological measurement, it seems obvious, should be as concerned with the positive, constructive, and life-enhancing forces within the personality as with the negative, destructive, and life-limiting. Likewise with respect to criteria: Measurement should be concerned with problems such as crime, psychiatric breakdown, accidents, and suicide, but it should also be concerned with happier outcomes such as achievement, creativity, and the attainment of goals.

Measures of aptitude and ability have an intrinsic relevance to these latter criteria, and studies using such instruments have typically been concerned with favorable outcomes of one sort or another. Measures of personality, on the other hand, have more often been addressed to dimensions and factors of psychopathology, and studies using these instruments have tended to deal with distress and disturbance.

However, some (perhaps many) psychologists would argue that the positive side of behavior is simply the absence or minimal presence of factors such as are assessed by standard clinical tools. For example, a great many studies of scholastic achievement have used clinical measures of personological malfunctioning, and this fact may attest to the belief just cited. Unfortunately, these studies

have yielded little if anything in the way of predictive validity (see Henry, 1950, p. 451).

In contrast, a persuasive case can be made for "dimensionalizing" the positive aspect of personality. The belief here is that constructive achievement is not merely a consequence of the absence of negative or restraining factors, but an outcome determined by forces and variables which must be defined and measured in their own right.

The California Psychological Inventory (CPI) (Gough, 1957) was introduced in the hope of providing such measurement. The variables chosen for scaling are conceived of as "folk concepts," dimensions of personality which arise organically out of social living and hence to be found everywhere, in all societies and cultures. The general theory of the CPI is sketched elsewhere (see, e.g., Gough, 1956), but the emphasis of the test upon interpersonal traits having broad relevance to personal and social adequacy should nonetheless be noted.

Scholastic achievement is one of the specific criteria which should be predictable from the CPI. Two of the inventory's scales deal explicitly with achievement motivation. The first of these, "Achievement via Conformance" or "Ac" (Gough, 1953c), attempts to calibrate that aspect of the achievement motive in which self-discipline, acceptance of rules, and convergent thinking predominate. The second, "Achievement via Independence" or "Ai" (Gough, 1953a), seeks to reflect that form of achievement motivation in which independ-

¹ Data for this study were gathered while the author was a Fulbright fellow at the Institute of Psychology, University of Florence, Italy. Analyses were supported by a 5-year basic research grant from the Ford Foundation, 1957-1962.

ence, the creation of method rather than observance of method, and divergent thinking are accentuated. Ac tends to yield its highest correlations in occupational and educational settings where conformance is a desirable attribute, and Ai in settings putting a greater emphasis on individuality and creative innovation.

Studies with the CPI are, in fact, beginning to show encouraging results in forecasting under- and overachievement (Butler, 1957; Fink, 1962), academic performance in high school (Gill, 1961), differential achievement among highly talented students (Holland, 1959; Lessinger & Martinson, 1961; Pierce, 1961), and achievement in particular subjects such as mathematics (Keimowitz & Ansbacher, 1960) and clinical psychology (Rosenberg et al., 1962). The folk concept perspective of the CPI and its intended relevance to social behavior everywhere suggest that these results should be checked by cross-cultural studies of the achievement motive. This paper reports on one such undertaking, a study of academic achievement at the high school (liceo) level in Italy.

METHOD

Subjects Ss in the study were students from four Italian high schools: two schools in Como, one in Florence, and one in Naples.² Northern, central, and southern regions of Italy were therefore represented in the sample. The total sample consisted of 204 males and 137 females, with subtotals by school for males of 59, 46, 33, and 66, and for females of 47, 13, 30, and 47.

Two tests were used in the study. The first was the CPI³ and the second was the "dominoes" test or D 48.⁴ The latter is a nonverbal test of general

² The help of the following persons in conducting this testing is gratefully acknowledged: Como—Leonardo Ancona and Anna Riva; Florence—Alberto Marzi and Giovanna Cannata; Naples—Gustavo Iacono, Maria Sbandi, and Giulia Villone.

³ An Italian translation of the CPI was used. Items in the Italian version are, with two or three exceptions, the same as in the English version, and scoring of all scales is identical. Many persons contributed to the Italian translation, but special mention must be made of the invaluable efforts of Francesca Morino Abbele. The Italian edition of the CPI is published by the Organizzazioni Speciali, via R. Franchi 5, Florence, Italy.

⁴ In Italy, nonverbal tests such as the D 48 and Raven's *Matrices* are generally favored for group testing. Because of this, and because of the unavail-

TABLE 1
MEANS AND STANDARD DEVIATIONS ON THE CALIFORNIA PSYCHOLOGICAL INVENTORY AND D 48
TEST FOR STUDENTS FROM FOUR ITALIAN HIGH SCHOOLS

Test	Males ^a		Females ^b	
	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>
CPI				
Do	25.44	5.44	23.18	5.49
Cs	16.90	3.59	15.28	4.07
Sy	21.37	4.88	19.46	5.16
Sp	31.54	6.56	29.42	7.00
Sa	20.05	3.74	19.18	3.64
Wb	32.56	5.67	30.28	6.08
Re	27.12	5.12	27.55	5.21
So	33.64	5.36	34.41	5.60
Sc	24.97	7.98	24.55	7.45
To	17.67	4.89	17.12	4.48
Gi	17.88	6.06	16.18	5.27
Cm	21.88	3.13	22.46	2.49
Ac	22.56	4.73	22.04	4.23
Ai	15.45	3.81	15.55	3.33
Ie	33.28	4.96	33.19	4.99
Py	9.40	2.70	8.97	2.62
Fx	8.03	3.67	8.26	3.56
Fe	16.70	3.40	22.07	3.52
D 48	29.72	4.68	27.03	5.91

^a *N* = 204.
^b *N* = 137.

intelligence first introduced by Anstey and Illing (see Gough & Domino, 1963) in England, and later published on the Continent (see Ferracuti & Rizzo, 1959; Pasquasy & Doutrepont, 1956). The D 48 is widely used in Europe, and is believed to be an excellent index of the *g* factor. Testing with the CPI and D 48 was conducted in the late fall and early winter, 1958–59.

Criteria for the study were derived from the grades ("scrutini") received in June 1959, covering the school year in which testing was done. The Italian curriculum for the high school is more or less standardized, but certain differences do exist between classical and scientific curricula. It might be informative simply to list the subjects at each school used to determine the grade average: Como, School A—Italian, Latin, Greek, history, philosophy, mathematics, history of art, physical education, and conduct; Como, School B—Italian, Latin, English, history, philosophy, mathematics, physics, science, design, physical education, and conduct; Florence—Italian, Latin, Greek, history, philosophy, mathematics, physics, science, art, physical education, and conduct; Naples—Italian, Latin, Greek, history, philosophy, mathemat-

ability of a well-standardized verbal group test, the decision was made to use the D 48.

ics, physics, science, history of art, physical education, and conduct.

For purposes of analysis, the Ss were first grouped into subsamples, by sex, two original samples consisting of students from School B at Como, Florence, and Naples, and cross-validating samples comprised of students from School A at Como. Grade averages for each school were converted to standard scores (mean of 50 and standard deviation of 10) before constituting the original and cross-validating samples. Analyses of the 19 test scores against the criterion of grades were then conducted.

RESULTS

Table 1 presents means and standard deviations on the 19 test scores for the total samples of 204 males and 137 females. The averages for the CPI are very similar to those found for high school students in the United States (Gough, 1957, pp. 34-35). Differences between means of three or more points are

found only five times out of 36 comparisons: Italian males are lower than American males on the Cm (Communality) scale, and Italian females are lower than American females on Wb (Sense of Well-Being), So (Socialization), Sc (Self-Control), and Cm (Communality).

Comparative American data for the D 48 test do not exist, but the Italian means of 29.72 (males) and 27.03 (females) are similar to those reported in other European studies (cf. Gough & Domino, 1963). Incidentally, the sex difference in the present samples, although statistically significant, is an exception to the general finding; in most studies the D 48 has not shown any difference between average scores of males and females.

Analytic statistics are given in Table 2 for the original and cross-validating samples. For males in the original sample, 12 of the 18 CPI

TABLE 2
CORRELATIONS OF TEST SCORES WITH GRADES IN THE ORIGINAL AND CROSS-VALIDATING
SAMPLES OF ITALIAN STUDENTS

Test variable	Male sample		Female sample	
	Original (<i>N</i> = 145)	Cross- validating (<i>N</i> = 59)	Original (<i>N</i> = 90)	Cross- validating (<i>N</i> = 47)
CPI				
Do	.18*	.32*	.16	.25
Cs	.02	.06	.28**	.03
Sy	-.02	.05	.13	-.03
Sp	-.12	-.05	.16	-.26
Sa	-.05	.07	.12	.08
Wb	.18*	.23	.27*	.41**
Re	.38**	.22	.17	.38**
So	.27**	.23	.08	.24
Sc	.23*	.06	.21*	.32*
To	.17*	.21	.37**	.29*
Gi	.19*	.08	.27**	.29*
Cm	.34**	.08	.08	.32**
Ac	.32**	.31*	.28**	.43**
Ai	.35**	.37**	.37**	.15
Ie	.28**	.24	.24*	.28*
Py	.15	.01	.21*	.07
Fx	-.33**	-.11	-.06	-.37**
Fe	.14	.26*	-.05	.00
CPI equations				
M—preliminary	.49**	.42**	—	—
F—preliminary	—	—	.40**	.37**
D 48	.36**	.07	.34**	-.04

* $p < .05$.

** $p < .01$.

TABLE 3

CORRELATIONS OF TEST SCORES WITH GRADES IN THE
TOTAL SAMPLES OF ITALIAN STUDENTS

Variable	Male ^a	Female ^b
CPI scales		
Do	.22**	.19*
Cs	.03	.19*
Sy	.00	.07
Sp	-.10	.00
Sa	-.02	.11
Wb	.19*	.32**
Re	.33**	.25**
So	.25**	.13
Sc	.19**	.25**
To	.18**	.34**
Gi	.17*	.28**
Cm	.27**	.17*
Ac	.32**	.33**
Ai	.35**	.29**
Ie	.26**	.25**
Py	.11	.16
Fx	-.25**	-.18*
Fe	.17*	-.03
CPI equations		
M—final	.46**	.40**
F—final	.45**	.43**
D 48 test	.28**	.07

^a *N* = 204.^b *N* = 137.* *p* < .05.** *p* < .01.

scales revealed significant ($p < .05$) correlations with the criterion; for females, nine of the 18 scales yielded significant coefficients. For both sexes the two measures of achievement motivation (Ac and Ai) gave significant values. The D 48 test also correlated significantly with grades in the original samples.

In order to maximize the predictive efficiency of the CPI two stepwise multiple-regression analyses were undertaken (one for each sex).⁵ The equation for males was as follows:

$$M\text{—preliminary} = 35.63 - .29Gi + .02Ac \\ + .90Ai + .41Ie - 1.17 Fx \quad [1]$$

In the original sample on which it was developed this equation correlated +.49 with the

criterion of grade average. When cross-validated on the sample of 59 males from School A at Como it gave a correlation of +.42, significant beyond the .01 level.

It might be noted that in both samples the predictive validity of the test of ability (D 48) was less than that of the CPI equation; however, it might well be that a verbal test of ability would have fared better as a predictor of scholastic performance.

The variables in the CPI equation should also be remarked. All three of the scales (Ac, Ai, and Ie) from the third cluster of the CPI profile sheet—described as “measures of achievement potential and intellectual efficiency” (Gough, 1957, p. 7)—are included in this predictive equation with positive weights. This finding would seem to offer support for the original specification of the cluster.⁶ The Gi (Good Impression) scale, in spite of an initial positive correlation with the criterion, takes on a negative loading in the equation, and the other component is the negatively weighted score on Fx (Flexibility). The psychological picture suggested by inspection of this combination of variables is of a highly motivated, intellectually efficient, and adaptive student who nonetheless adheres strictly to the rules and is careful not to overstate his desirable views of himself.

The equation yielded in the analysis for the 90 females in the original sample was this:

$$F\text{—preliminary} = 33.56 + .22Ac \\ + 1.01Ai - .41Fx \quad [2]$$

This equation correlated +.40 with grades in the original sample, and on cross-validation in the sample of 47 students from School A at

⁶ In several reports on the CPI it has been said that the grouping of scales on the profile sheet was done by subjective inspection only. This statement is incorrect. Prior to the publication of the *Manual* in 1957, five factor and/or cluster analyses were conducted (three on samples of males, two on samples of females), and the clusters on the profile sheet were defined so as to emphasize the common pattern observed among these analyses. Clusters I and II on the profile sheet are largely factorially determined. Cluster III (the achievement cluster) was defined for the convenience of counselors and others who might find it useful to have the achievement indices grouped together. Cluster IV includes variables having only slight loadings on the earlier factors.

⁵ These analyses were conducted at the University of California, Berkeley, Computing Center and were monitored by Karl Scheibe and Quintin Welch.

Como gave a coefficient of $+ .37$, significant beyond the $.01$ level. As before, the two coefficients exceed those given by the test of intellectual ability.

The CPI equation for women again contains both of the achievement-motivation scales (Ac and Ai), but Ie (Intellectual Efficiency) does not appear. The third variable is the Flexibility scale. The psychological pattern suggested by the equation of three scales is once more that of a motivated, adaptive, conscientious student who cathects form and structure and who, although capable of independent and individual endeavor, nonetheless abides resolutely by the stipulations of academic authority.

The confirmation of both equations in cross-validation led to the decision to seek predictive equations of maximum validity, using the total samples. New analyses were therefore conducted, using the full samples of 204 males and 137 females. Table 3 presents the findings.

For males, 13 of the 18 scales of the CPI correlated significantly ($p < .05$) with grades, and for females 12 of the 18 scales reached this same level. Ac and Ai were both significant beyond the $.01$ level for both sexes. The D 48 test gave a significant correlation for males ($r = +.28$), but an insignificant coefficient ($r = +.07$) for females.

Stepwise multiple-regression analyses were again conducted, by sex. For males, the optimum equation included seven scales of the CPI. However, a much simpler equation using only three scales gave approximately equal results and was accordingly chosen for presentation:

$$M\text{---}final = 35.92 + .92Ai \\ + .18Ie - .81Fx \quad [3]$$

This equation correlated $+ .46$ with grades in the sample of 204 males on which it was developed. When applied to the sample of 137 females it correlated $+ .40$ with grades. This second correlation may be considered a cross-validational value.

M—final gave a predicted mean of 49.59, standard deviation of 4.83, for the 204 males, and for the 137 females a predicted mean of 49.47, standard deviation of 3.97. Its constant of 35.92 thus led to a slight underestimation

of grade averages, as the actual values were $M = 49.62$, $SD = 10.33$ for males, and $M = 50.72$, $SD = 9.17$ for females.

The equation developed on the female sample was this:

$$F\text{---}final = 34.77 + .33Ac \\ + .85Ai - .56Fx \quad [4]$$

Scores calculated on the basis of this equation correlated $+ .43$ with grades in the female sample, and $+ .45$ in the cross-validating sample of males. F—final somewhat overestimated grades, yielding a mean of 51.15, $SD = 5.51$ for males, and a mean of 51.12, $SD = 4.27$ for females. This could easily be corrected by reducing the constant in the equation from 34.77 to 33.63.

The correlation between the scores on M—final and F—final was $+ .97$ in the sample of 204 males, and $+ .95$ in the sample of 137 females. The similarity is obviously great, suggesting that it makes little difference which equation is used. F—final appears to be just slightly more valid as a predictor of grades, and so in future studies the recommendation is made that this equation be used to compute predicted grades for both sexes.

In spite of the fact that the ability test (D 48) did not fare too well as a predictor of grades in this study, a measure of ability might do better in another investigation. Hence, the correlations between M—final and F—final and D 48 are of interest. For males, D 48 correlated $+ .21$ with M—final and $+ .20$ with F—final; for females the corresponding values were $+ .17$ and $+ .19$. The CPI equations, one may conclude, are functioning fairly independently of general ability and may be presumed to be tapping factors of a "nonintellectual" variety.

DISCUSSION AND INTERPRETATION

Too frequently, studies of positive outcomes and of personal achievement have utilized personal measures of distress, anxiety, and clinical malfunctioning. The theoretical problem rests on one's conception of the kind of variables needed to explicate and forecast constructive attainment. The author's view is that personal adequacy is not just an absence of pathology and inner turbulence, but a pattern-

ing and combination of factors that must be identified and measured in their own right.

The history of failure in predicting scholastic achievement from personological variables is seen as stemming, in part, from this attempt to forecast positive outcomes from measures of disturbance. The California Psychological Inventory, deriving from a theoretical perspective emphasizing positive, achievement-oriented dimensions of personality arising ineluctably out of social interaction, represents a contrasting approach.

Studies of scholastic achievement with the CPI in different settings and with different age groups have, as noted above, yielded preliminary but encouraging results. The scales of the inventory are intended to have universal relevance, i.e., to assess folk concepts of personality, and it is therefore appropriate to seek cross-cultural validation. In particular, with respect to academic performance, the two CPI measures of achievement motivation merit such investigation.

The Ac (Achievement via Conformance) and Ai (Achievement via Independence) scales did, in fact, correlate significantly with the scholastic performance of Italian males and females. The evidence, i.e., is in support of the cross-cultural relevance and validity of these two measures of achievement motivation. Furthermore, both scales are included, with positive weightings, in the optimum regression equation for predicting GPA from the scales of the inventory. Although the study is limited to criteria of academic achievement, and must therefore be interpreted with caution, its findings may be said to favor the theory that folk dimensions of personality have universal relevance and that measures of these dimensions will partake of this universality; vice versa, its findings are at least to some extent unfavorable to the contrary view, widespread in present-day psychometrics, that personality variables and measures are culturally bound and specific.

REFERENCES

- BUTLER, J. J. Differential factors in the self-concepts of overachieving, underachieving, and expected-achieving adolescents. Unpublished doctoral dissertation, University of Southern California, 1957.
- FERRACUTI, F., & RIZZO, G. B. Studio sul test D-48 applicato ad una popolazione italiana di livello scolastico superiore. (A study of the D 48 test as applied to an Italian population of superior educational level.) *Boll. psicol. sociol. appl.*, 1959, No. 31-36, 77-83.
- FINK, M. B. Objectification of data used in under-achievement self-concept study. *Calif. J. educ. Res.*, 1962, 13, 105-112.
- GILL, LOIS J. Some non-intellectual factors related to the academic achievement of Spanish-American secondary school students. Unpublished doctoral dissertation, University of Denver, 1961.
- GOUGH, H. G. The construction of a personality scale to predict scholastic achievement. *J. appl. Psychol.*, 1953, 37, 361-366. (a)
- GOUGH, H. G. A nonintellectual intelligence test. *J. consult. Psychol.*, 1953, 17, 242-246. (b)
- GOUGH, H. G. What determines the academic achievement of high school students? *J. educ. Res.*, 1953, 46, 321-331. (c)
- GOUGH, H. G. Potential uses of personality scales in schools and colleges. In *Fifth annual western regional conference on testing problems—1956*. Berkeley, Calif.: Educational Testing Service, 1956. Pp. 3-20.
- GOUGH, H. G. *Manual for the California Psychological Inventory*. Palo Alto, Calif.: Consulting Psychologists Press, 1957.
- GOUGH, H. G., & DOMINO, G. The D 48 test as a measure of general ability among grade school children. *J. consult. Psychol.*, 1963, 27, 344-349.
- HENRY, E. R. Predicting success in college and university. In D. H. Fryer & E. R. Henry (Eds.), *Handbook of applied psychology*. Vol. 2. New York: Holt, Rinehart & Winston, 1950. Pp. 449-453.
- HOLLAND, J. L. The prediction of college grades from the California Psychological Inventory and the Scholastic Aptitude Test. *J. educ. Psychol.*, 1959, 50, 135-142.
- KEIMOWITZ, R. I., & ANSBACHER, H. D. Personality and achievement in mathematics. *J. indiv. Psychol.*, 1960, 16, 84-87.
- LESSINGER, L. M., & MARTINSON, RUTH A. The use of the California Psychological Inventory with gifted pupils. *Personnel Guid. J.*, 1961, 39, 572-575.
- PASQUASY, R., & DOUTREPONT, G. Le test des dominos (D.48). (The dominoes test (D 48)). *Bull. Orient. scol. Profess.*, 1956, 5, 20-34.
- PIERCE, J. V. Personality and achievement among able high school boys. *J. indiv. Psychol.*, 1961, 17, 102-107.
- ROSENBERG, L. A., MCHENRY, T. B., ROSENBERG, ANNA M., & NICHOLS, R. C. The prediction of academic achievement with the California Psychological Inventory. *J. appl. Psychol.*, 1962, 46, 265-268.

(Received June 17, 1963)

A "CONTINGENT-ITEM" METHOD FOR CONSTRUCTING A SHORT PERSONALITY QUESTIONNAIRE¹

FRANK B. McMAHON, JR.

AND

RAYMOND G. HUNT

Washington University, St. Louis

State University of New York at Buffalo

A brief, face-valid personality test was devised to be used as a limited range screening mode to locate general problem areas and the degree to which an individual is operating under "stress" in his daily life. The test uses a novel double-question method in which the 2nd question elaborates the answer to the 1st and is contingent upon it. This format is intended to encourage candidness on the part of the respondent and to enhance the power of the individual item. Criterion analyses and reliability indicate that the test and/or item form may prove useful in personality assessment.

This report summarizes limited and preliminary explorations into the merits of a unique item form intended to enhance measurement in brief personality questionnaires. As a generality, it seemed desirable to develop a searching item form that would, at the same time, facilitate candidness on the part of the test taker so that problems deleterious to his everyday functioning could be effectively measured. To accomplish this, while avoiding certain "pitfalls" of the short test, discussed by Thorndike (1949) and by Cronbach (1960), a brief series of items was devised in couplets such that each "basic" item was followed by a second, "contingent" item referring to the former and amplifying the implications of response to it.

This is illustrated by the following item couplet taken from the 48-item demonstration scale used:

*I have periods of such great restlessness
that I cannot sit long in a chair.*

The subject (S) responds, in forced-choice fashion, either "True" or "False." If, on this item, he responds "True," he is then asked to respond "True" or "False" to the following contingent statement:

*Most of the time I feel like a rubber band
stretched tight.*

Use of this format, with its emphasis upon face validity, is intended to allow the test

taker to feel he is in control of the situation inasmuch as he can explain (partly, at least) what he "means" by a response. This, it is further posited, will tend to facilitate candidness on his part because he has greater opportunity to both tell the truth and feel less fear that what he says will be misinterpreted.

In sum then, we may describe the principal advantages ascribed to the "48-item test" as it has so far developed with this format as follows:

1. Administration time is minimal—about 10 to 20 minutes.
2. Test items are varied enough to cover most problem areas affording the tester-counselor the opportunity to do a quick "content analysis" before talking with the client and providing a potential starting point for counseling.
3. The test format is designed mainly to provide more "powerful" individual items and to encourage frankness from the S.

The test, in its present stage of development, is designed mainly as a general screening device to identify mild maladjustments, especially in counseling situations where group forms of tests are administered in batteries. Its usefulness in other contexts is still unknown.

DEVELOPMENT OF THE PRESENT SCALE

First, from the 550 items in the MMPI pool, 50 were selected to fit four initial criteria: (a) items were to be representative of major areas of psychological functioning (so-

¹ This paper is based upon a master's thesis submitted by F. B. McMahon to the Graduate Board of Washington University, St. Louis.

TABLE 1

POINT-BISERIAL CORRELATIONS FOR MMPI AND CONTINGENT-ITEM TEST IN THREE STUDIES USING CRITERION OF SUBJECTS' PERSONAL FEELINGS OF NEEDS FOR COUNSELING

Study	Contingent-item test	MMPI items alone
I ($p = .64$, $N = 59$)	.56*	.35*
II ($p = .73$, $N = 101$)	.52*	.25*
III ($p = .57$, $N = 23$)	.59*	.28

* $p \leq .01$.

cial life, home life, sexuality, fantasy life); (b) they were to be indicative of clear, but undifferentiated, pathology (i.e., face validity was stressed); (c) items needed to be amenable to the "if-true" format; (d) items were *not* to appear "tricky" or "oblique."

A contingent unit was then written for each of these so as to "probe" the first unit and it formed the second part of a couplet. Finally the test was administered to groups of college students and to vocational counseling clients in a series of six separate studies.

After examination of 185 test administrations the original 50-item test was reduced to 45. Three new items were later added, each of which was "original," i.e., not drawn from the MMPI. We might note that use of MMPI items is not required by the procedure. Their use was purely a matter of convenience and their careful construction. Actually these items are now being suitably revised.²

The 48 items constituting the test were designed to tap feelings of (a) loneliness and/or (b) frustration and/or (c) tension in the person's life.

Empirical Analyses. As noted, six separate studies were performed to provide preliminary estimates of the effectiveness of the contingent-item form and of the validity and reliability of a brief, general personality inventory using such items (the 48-item test). The procedures and results of these investigations are summarized below. A more detailed treat-

ment of these and of the rationale for the item form may be found in McMahon (1961).

The first three studies were designed chiefly to investigate the premise that a contingent-item form of test construction would discriminate more effectively than would a more conventional form between persons experiencing psychological "problems" and those who are not. A total of 183 students, ranging in age from 17 to 40 years, from various psychology courses served as Ss. Of these 117 were male. Procedures were approximately the same in all three studies. To wit, a form of the 48-item test was given in a group administration. The Ss were told the test was part of a research project and were encouraged to write comments on the test blank. A general discussion followed testing.

As a very tentative criterion each *S* was asked to answer "True" or "False" to a form of the following question appearing on a separate sheet appended to the test blank: "I feel that my personal-life situation is such that I am in need of some form of psychological assistance (therapy, counseling, etc.) with my personal problems." Though by no means an ideal criterion it was assumed that *S* should have a fair idea of his need for such assistance and that such felt needs would, in some useful degree, index underlying personality maladjustments.

The results of these studies are presented in Table 1 in the form of pairs of point-biserial correlations. For these correlations a 48-item test score was obtained by tallying for each *S* the frequency with which contingent items in a couplet were checked so as to indicate "pathology."³ The cumulative or total score for *S* on the 48-item test is thus the simple sum of contingent items checked in the keyed direction.

A separate MMPI score was obtained by taking the total number of initial units in a couplet answered pathologically (as scored by the MMPI key) for each *S*. These two test scores were independently correlated with the dichotomous "counseling" criterion.

It can be seen from Table 1 that the 48-

² A version of the test, of more recent vintage, will be published shortly by Western Psychological Services.

³ The term "pathological" as used here means that *S*'s response to an item differs from what would "rationally" be expected from one with "no problems."

item test fares quite well, both in its own right and relative to the MMPI score alone. Also, the 48-item correlations are notably stable across studies. Pooled frequency distributions of 48-item and MMPI cumulative scores show that 48-item scores cluster around 0, 1 and 2 whereas the MMPI cumulative scores show a much wider scatter.

Further Criterion Measures. Two other studies were instituted mainly to explore additional more independent criterion measures. The first used as Ss 14 college students (of whom five were male) enrolled in a general psychology course. These students worked together closely and were mutually well acquainted. Each S took the 48-item test and then ranked his classmates in descending order of their need for psychological counseling. These peer nominations were then compared with categorized cumulative 48-item scores.

For this analysis a cutting score of four on the test was set as indicative of pathology. This score was arrived at by noting that in preceding studies nearly all Ss reporting "not-needing-counseling" scored four or less.

Of the 14 Ss rated only three were consistently selected as most "disturbed." The remaining Ss all earned low *average* ranks. None of the latter was selected as pathological by the 48-item test while two of the former were. Hence the test agreed with peer nominations in 13 of 14 instances, producing no false positives and only one miss.

The last "validity" study was undertaken at the Washington University Adult Counseling Center to determine three things: (a) the feasibility of using the test under actual counseling conditions, (b) the correlation of the test with blind counselors' ratings of clients, and (c) the correlation of the test with other, longer personality measures.

Twenty-five vocational counseling clients served as Ss, the first 25 in order of appearance at the Center (data from two of these Ss had ultimately to be discarded). Their age range was from 27 to 41 years and five were male. The 48-item test was administered as part of a regular test battery. Counselors were not apprised of test results until they had "voted" on whether or not the client was char-

TABLE 2
CORRELATIONS BETWEEN CONTINGENT-ITEM TEST SCORES ABOVE AND BELOW A CUTTING SCORE OF FOUR AND THREE INDEPENDENT CRITERIA OF ADJUSTMENT

Contingent-item test	Criterion	
	Counselor's ratings	
	Yes	No
Yes	7	2
No	0	14
	$\phi = .82; \chi^2 = 15.41^*$	
	Guilford-Zimmerman	
	Yes	No
Yes	7	1
No	2	11
	$\phi = .76; \chi^2 = 12.18^*$	
	Full-scale MMPI	
	Yes	No
Yes	9	0
No	1	9
	$\phi = .90; \chi^2 = 15.39^*$	

* $p < .01$.

acterized by a significant degree of maladjustment, based upon their observations in a personal interview. During testing each S received either or both the full-scale MMPI or Guilford-Zimmerman Temperament Survey.

The fourfold-point (ϕ) correlation between counselors' ratings of the clients' adjustment and similar selections based on the 48-item test (again using a cutting score of four) is presented in the upper section of Table 2. Point-biserial correlations were also computed between the MMPI question units alone and the contingent units, using the counselors' ratings as the criterion. The values of these were .36 and .82, respectively ($p = .76$, $N = 21$), results consistent with those reported earlier.

A ϕ coefficient of .76 was obtained between the 48-item test and the Guilford-Zimmerman and a similar correlation of .90 with the full MMPI (see Table 2). For both of these correlations the cutting score of four on the 48-item test was taken as indicating maladjustment. On the Guilford-Zimmerman,

a "peak" (10% or lower) on any one or more temperament categories served the same purpose. The MMPI maladjustment criterion used was a peak (adjusted T score of 70 or over) on any one or more of the "clinical" scales.

We might finally point out that the counselors generally found the 48-item test to be useful and "clinically" helpful in its intended manner. In fact the test (with revised items) is now a regular component of the Counselling Center's battery.

Reliability. A preliminary estimate of the test-retest reliability of the 48-item test was obtained from 42 psychology students ranging in age from 20 to 43 years. Twenty-two of these were male. The test was given in a group administration and the test-retest interval was about 3 weeks. The obtained correlation was .80.

We might note that obtaining reliability estimates for a test using the contingent-item format does pose something of a problem. While the correlation obtained is adequate for a test of the length of the 48-item test, it should be borne in mind that answers to the contingent unit of a couplet depend upon prior responses to the first unit. It was found that on the second administration changes in the latter precluded response to the former a total of 17 times. Still, using the provisional cutting score, only three cases would have shifted from a "maladjusted" to an "adjusted" category or vice versa. Therefore, it would seem that the reliability estimate obtained indicates good stability, especially when compared with longer tests.

The Problem of Dissembling. While the contingent-item form is intended to encourage candidness on the part of the S, and apparently does in large measure (most of our Ss preferred it to more conventional forms), it is still possible, obviously, for a respondent to be less than frank. Therefore, some kind of validity check is useful to identify such instances. No detailed work has been done on this, but the 48-item test does appear to have

within it a rough measure of the test taker's honesty in replying to it.

Of the approximately 300 persons who have taken the test in initial studies, only three have checked all of the initial units in the couplet items "favorably" toward themselves. Any test in which *all* such items are favorable would seem, therefore, to be immediately suspect. The same is not true of the second units, however. Here scores should generally be low (at least in populations like those tested so far), with an average of about one (1) question checked *unfavorably*.

DISCUSSION AND CONCLUSIONS

All of the findings reported above, while obviously modest, do appear to support the premise motivating the studies outlined, i.e., that the contingent item offers a provocative device for the construction of shorter, more searching personality-measurement instruments. More particularly it would appear from the data reported that the 48-item test using this question form gives promise of usefulness and so warrants further work.

Further study should include more detailed item analysis and selection, more extended criterion studies, analysis of the prospects for attaching diagnostic and/or prognostic rubrics to patterns of response, and comparison of test data with interview data. It should prove especially useful to investigate the test and/or item form in other contexts and with other kinds of subjects, e.g., hospital populations, so as to determine its functional limitations. Some of these studies, of course, are in progress.

REFERENCES

- CRONBACH, L. J. *Essentials of psychological testing*. New York: Harper, 1960.
- McMAHON, F. B. The "forty-eight item test": A short personality test for use in the counselling situation. Unpublished master's thesis, Washington University, St. Louis, 1961.
- THORNDIKE, R. L. *Personnel selection*. New York: Wiley, 1949.

(Received June 21, 1963)

KENNETH E. CLARK, Editor
University of Rochester

Table of Contents

A Quantitative Analysis of Expressed Preferences for Compensation Plans.....	Lyle V. Jones and Thomas E. Jeffrey	201
An Industrial Use of Peer Ratings.....	Harry E. Roadman	211
Selecting Competent Raters.....	Llewellyn Wiley and William S. Jenkins	215
Prediction of Performance in Medical School from the California Psychological Inventory.....	Harrison G. Gough and Wallace B. Hall	218
The Effect of Participatory and Supervisory Leadership on Group Creativity.....	Lynn R. Anderson and Fred E. Fiedler	227
Scaling Preferences for Television Shows.....	Frank J. Dudek and Evelyn Thoman	237
The Factorial Structure of Selected Consumer Choice Parameters and Their Relationship to Personal Values.....	John R. Rizzo and J. C. Naylor	241
Tracking with a Differential Brightness Display: II. Peripheral Tracking.....	Stanley M. Moss	249
The Effect of Time Stress and the Elimination of Cue Information on the Display-Control Relationships of Moving Scale Instruments.....	A. B. Hill and C. Q. Large	255
Supervisor Perception of Work Group Morale.....	Thomas H. Jerdee	259
Knowledge of Performance as an Incentive in Repetitive, Monotonous Tasks.....	Alphonse Chapanis	263
An Experimental Study of the Relation between Nursing Care and Patient Welfare.:.....	J. Richard Simon and Wellborn R. Hudson	268

American Psychological Association

Consulting Editors

GEORGE E. BRIGGS, *Ohio State University*
MARVIN D. DUNNETTE, *University of Minnesota*
NORMAN FREDERIKSEN, *Educational Testing Service*
LEONARD D. GOODSTEIN, *University of Iowa*
EDWIN R. HENRY, *Peace Corps*
JOHN HOLLAND, *American College Testing Program*
CLIFFORD E. JURGENSEN, *Minneapolis Gas Company*

LAURENCE S. MCGAUGHRAN, *University of Houston*
QUINN MCNEMAR, *Stanford University*
HAROLD F. ROTHE, *Beloit Corporation*
THOMAS A. RYAN, *Cornell University*
ALEXANDER G. WESMAN, *Psychological Corporation*
CLARK L. WILSON, *Batten, Barton, Durstine, and Osborn, Incorporated*

This journal gives primary consideration to original investigations in any field of applied psychology except clinical psychology, although a descriptive or theoretical article may be accepted if it represents a special contribution in an applied field. Quantitative investigations of interest or value to psychologists working in the following broad fields will be considered: vocational and educational prognosis, diagnosis, and guidance at the secondary and college level; personnel research in business, industry, and government; engineering psychology; industrial working conditions; research on opinion and morale factors; job analysis and classification research; market and advertising research.

Manuscripts must be accompanied by an abstract of 100–120 words typed on a separate sheet of paper. The abstract should conform to the style of *Psychological Abstracts*. Detailed instructions for preparation of the abstracts appeared in the *American Psychologist* (1961, 16, 833), or they may be obtained from the Editor or from the APA Central Office.

Manuscripts should be addressed to the Editor

Dr. Kenneth E. Clark
College of Arts and Science
302 Morey Hall
University of Rochester
Rochester, N. Y. 14627

All manuscripts must be submitted in duplicate. Original figures are prepared for publication; duplicate figures may be photographs or pencil-drawn copies.

Manuscripts must conform to the style requirements described in the *Publication Manual of the American Psychological Association*.

The following policies govern reprinting of materials copyrighted in APA journals: (a) to require approval to reprint for tables and figures, and for text only if more than 500 words [total from one article] in length; (b) to continue approval to reprint, contingent on the author's approval, for articles reprinted in whole or in major part; (c) to negotiate where possible the dedication of royalties from commercial publishers to the American Psychological Foundation.

Journal of Applied Psychology

Published bimonthly by the
American Psychological Association
Prince and Lemon Sts., Lancaster, Pa. 17604
and 1333 Sixteenth St. N. W.
Washington, D. C. 20036

\$10.00 per volume

\$2.00 per issue

HELEN ORR
Managing Editor

ELIZABETH S. REED
Advertising Manager

FRANCES L. BREWER
Editorial Assistant

Subscriptions, orders, and business communications should be addressed to the American Psychological Association, 1333 Sixteenth St. N. W., Washington, D. C. 20036. Address changes must reach the subscription office by the tenth of the month to take effect the following month. Undelivered copies resulting from address changes will not be replaced; subscribers should notify the post office that they will guarantee second-class forwarding postage. Other claims for undelivered copies must be made within four months of publication.

Second class postage paid at Lancaster, Pa. and at additional mailing offices.

© 1964 by the American Psychological Association, Inc.

Journal of Applied Psychology

VOL. 48, No. 4

AUGUST 1964

A QUANTITATIVE ANALYSIS OF EXPRESSED PREFERENCES FOR COMPENSATION PLANS ¹

LYLE V. JONES AND THOMAS E. JEFFREY ²

University of North Carolina

The method of factorial paired comparisons is employed in 2 studies designed to evaluate employee preferences for alternative forms of job compensation. Explicitly considered are 4 compensation features with 2 levels each: weekly salary vs. hourly wage, use or nonuse of supervisory merit-ratings, inclusion or exclusion of a piece-incentive plan, and pay increase vs. no increase. A 2⁴ factorial design provides estimates for and tests of significance on preference scale values associated with each compensation "package," as well as for scale contrasts between the 2 levels of each separate compensation feature.

Within the field of psychology have been developed a set of methods designed to transform qualitative data—e.g., human attitudes, judgments, values, preferences—into quantitative form (e.g., Guilford, 1954; Thurstone, 1959; Torgerson, 1958). These methods have their origin in psychophysics, the attempt to establish functional laws relating "magnitude" of judgment to measurable physical characteristics of stimuli. The more pertinent work, beginning with several papers by L. L. Thurstone in 1927, provides models by which a quantitative scale may be established for "value" judgments of objects or events even when physical magnitude of such objects or events is not possible to define. To the extent found applicable to empirical data, these models provide means for assessing such intrinsically qualitative phenomena as consumer preference for competing commodities, attitudes towards political issues, or esthetic judgments pertain-

ing to the value of artistic accomplishments. Our aim in this paper is to present results from two studies in which a psychological scaling model is employed to evaluate employee preferences for alternative forms of job compensation.

THE MODEL

The method selected for the two studies is that of factorial paired comparisons, using an incomplete factorial design.³ An experimental situation is required whereby judges or subjects (*Ss*) provide pairwise responses to a set of stimuli by the following rules.

- (a) A set of *n* distinct stimuli is prepared.
- (b) On each trial in the experiment *S* is presented two objects from the set. He is required to choose the object in this pair which he prefers.
- (c) The objects which make up the pair for each trial are specified by the paired comparison design. In an incomplete design (such as used for the present studies) certain pairs are omitted.
- (d) Each *S* judges all pairs specified by the design.

¹ The authors are indebted to R. Darrell Bock for helpful advice regarding design and analysis, to J. Stacy Adams for suggestions pertinent to experimental design and for aid in experimental administration at Plant 1, to Stanley Nealy for supervision of data collection at Plant 2. The study was made possible by financial support from the Behavioral Research Service, General Electric Company.

² Now at the Army Personnel Research Office, Washington, D. C.

³ The detailed model, and the methods for its statistical evaluation, are presented by Bock and Jones (1963, Sec. 5.2).

(e) The response of *S* to a given stimulus pair is recorded by means of a formal variate, assigned a value of zero or one, depending upon whether the one or the other stimulus is preferred.

Assume that, prior to the experiment, stimuli have been classified in accordance with a factorial model; such a model is the basis for many analyses of variance designs. One such model, and that utilized in the two studies reported below, is a 2⁴ factorial design. Each stimulus may be represented in a single cell of a 2 × 2 × 2 × 2 table, where each of the four dimensions is associated with a qualitative characteristic of the stimuli, and each characteristic takes on either of two distinct states. A stimulus here may be represented symbolically by *X_{jk_{lm}}*, where *j*, *k*, *l*, *m* take on values one or two, depending upon whether the characteristics along the first, second, third, and fourth dimensions, respectively, are in the state labeled one or the state labeled two. Then the response of the *i*-th *S* to a pair of stimuli is assumed to be dependent upon the difference between affective processes associated with the two stimuli, each of the form

$$v_{jklm,i} = \mu + \alpha_j + \beta_k + \gamma_l + \delta_m + (\alpha\beta)_{jk} + (\alpha\gamma)_{jl} + \cdots + (\gamma\delta)_{lm} + \epsilon_{jklm,i}; \quad [1]$$

μ represents the population grand mean, averaged over all stimuli; α , β , γ , and δ are parameters representing main effects of the classification variables; the six first-order interactions are represented by the terms within parentheses; higher-order interactions are combined with error variance attributable to imperfect fit of the model⁴ into a residual term $\epsilon_{jklm,i}$, considered to be distributed normally over judges, with population mean zero and population variance σ_e^2 , assumed to be constant for all stimuli. Under these conditions, the difference between affective processes for any two stimuli also has a normal distribution with constant variance. The proportion of *Ss* who respond with preference for the first rather than the second of a pair of stimuli is thus related to the affective difference by the

⁴ Actually, only those higher-order interactions which fail to reach significance when compared with the error variance are then combined with the error variance; other higher-order interactions would then appear as explicit terms of the model specified by equation (1).

integral over the normal distribution,

$$p_{12} = \frac{1}{\sqrt{2\pi}} \int_{-(v_1-v_2)/\sigma}^{\infty} e^{-\frac{1}{2}z^2} dz. \quad [2]$$

An inverse normal transformation of *p*₁₂ thus yields an estimate for (v₁ − v₂)/σ.

A transformation so closely approximating the normal as typically to be indistinguishable from it on the basis of empirical data is the angular transformation,

$$p_{12} = \frac{1}{2} \int_{-(v_1-v_2)}^{\pi/2} \sin \left(\frac{\pi}{2} - z \right) dz, \quad [3]$$

from which v₁ − v₂ may be estimated from *p*₁₂ by the expression

$$y_{12} = v_1 - v_2 = 2\sin^{-1} \sqrt{p_{12}} - \pi/2. \quad [4]$$

An advantage of this transformation is that it tends to stabilize the variance of *y*₁₂ at 1/*N*, where *N* is the number of *Ss* who judge the stimulus pair 1, 2. The angular transformation, then, provides estimates for differences, *y*, between pairs of *v*'s, using tables of the arcsine function. This transformation is adopted for analysis of data from the experiments reported below.

With a 2⁴ factorial design, we have 16 distinct stimuli. If each stimulus were compared with each other stimulus, 120 stimulus pairs would be presented to *S* for evaluation. A balanced incomplete paired comparison design is available for 16 objects, in which each stimulus is compared with 3 other stimuli, requiring the presentation of 48 pairs of stimuli to each judge (McKeon, 1960).

EXPERIMENT 1

A total of 194 male employees of an electrical equipment plant located in the southeastern

TABLE 1
THE FACTORIAL CLASSIFICATION SCHEME

Classification	States	
	1	2
A	Hourly wage	Weekly wage
B	No merit incentive	Merit incentive
C	No piece-work incentive	Piece-work incentive
D	Lower pay rate	Higher pay rate

United States (Plant 1) served as Ss. They were tested in groups of from 11 to 21 Ss each. The Ss were asked to choose between alternative compensation plans, each of which could be characterized by a combination of features defined by the classification scheme of Table 1. Exact provisions of each plan were described to Ss in considerable detail before the comparative judgments of the 16 compensation plans were obtained. Stimuli were

presented by slide projection and were simultaneously read aloud by the experimenter. Two typical stimulus pairs are displayed in Figure 1.

For each classification in Table 1, the condition denoted State 1 is a feature of compensation existing in the plant at the time of this investigation. For each classification, State 2 represents an innovation; for Classifications A, B, and C the conditions represented by

ITEM 6

Plan (2121)

a) weekly salary
b) lower pay class- \$77.44 per week
c) no merit rating
d) additional 72 cents per hundred pieces over quota

L

Plan (2112)

a) weekly salary
b) higher pay class- \$79.38 per week
c) no merit rating
d) no piece incentive

R

ITEM 41

Plan (1111)

a) hourly wage
b) lower pay class- \$ 2.00 per hour
c) no merit rating
d) no piece incentive

L

Plan (2212)

a) weekly salary
b) higher pay class- \$79.38 per week
c) merit rating - pay may be as low as \$73.57 per week or as high as \$85.18 per week
d) no piece incentive

R

FIG. 1. Two illustrative items.

TABLE 2—PROPORTIONS WITH WHICH ROW PLANS ARE PREFERRED TO COLUMN PLANS, EXPERIMENT 1

Factors ABCD	1111	1112	1121	1122	1211	1212	1221	1222	2111	2112	2121	2122	2211	2212	2221	2222
1111																
1112							.155	.170					.438	.247	.247	
1121														.366	.366	
1122		.526												.335		
1211	.701												.572			.273
1212							.526									.216
1221																
1222																
2111					.871				.680		.521		.680			
2112									.696			.304		.443		
2121			.479								.691	.098				
2122	.747				.325			.237								
2211					.598											.119
2212			.510		.655			.505	.660			.459			.536	
2221	.825	.691				.546				.799	.649					
2222			.727													

States 1 and 2 were designed so as to be of equal cost to the company. For weekly salary, Classification A, the benefits emphasized were paid sick leave, personal leave time, longer paid vacation, and other minor distinctions from the hourly pay basis. Equivalent to \$2.00 per hour (the lower pay rate) in expected cost to the company is a weekly salary of \$77.44, equivalent to \$2.05 per hour (the higher pay rate) is \$79.38 per week. For merit incentive, each worker is rated for excellence by his supervisor into one of seven rating categories. An "average" rating entitles him to the standard rate. Ratings in other categories lead to pay differing from the standard by five cents per hour for each departure of one category from the average. Under the piece-incentive state, Classification C, workers are paid an added amount of about $\frac{3}{4}$ cent per piece for each piece in excess of their quota. With no piece incentive, they are expected to produce the quota, but receive no added pay for exceeding it. The piece unit was selected from a standard job at the plant, familiar to all Ss.

In Table 2 appears the proportion of the 194 Ss who expressed preference for each compensation plan designated by the row code to the plan designated by the column code, for the 48 plans included in the incomplete paired-comparison design. The angular deviates y_{jk} , transformed by (4) from the observed proportions, are shown in Table 3.

In a balanced incomplete paired-comparison design, the affective value of any stimulus v_j may be estimated from the expression

$$v_j = \frac{r}{n\lambda} (y_{j\cdot} + \sum_h v_{hj}y_{h\cdot}/r), \tag{5}$$

where r is the number of other stimuli with which each stimulus is paired, n is the total number of stimuli, λ is the number of stimuli commonly paired with any two selected stimuli, $y_{j\cdot} = \sum_k y_{jk}$, and $v_{hj} = 1$ if the h, j pair is included in the design, and is zero, otherwise. In the present study, $r = 6$, $n = 16$, $\lambda = 2$, so that (5) becomes

$$v_j = \frac{6}{(16)(2)} (y_{j\cdot} + \sum_h v_{hj}y_{h\cdot}/6). \tag{6}$$

TABLE 3
ANGULAR DEVIATES AND ESTIMATED AFFECTIVE VALUES, EXPERIMENT 1

Factors ABCD	1111	1112	1121	1122	1211	1212	1221	1222	2111	2112	2121	2122	2211	2212	2221	2222
1111																
1112																
1121																
1122		0520	-0520				-3152	-7208			-0921	-5167	-1243	-5305	-5305	-7076
1211	7615												-0200	-3920	-2713	-4713
1212	4137												1445	-3363	0520	-4713
1221																-6041
1222																
2111																
2112																
2121																
2122																
2211	5167															
2212	5305															
2221	5305															
2222	7076															
y_j	-3.4605	-1.6525	-1.0488	.8963	-1.6734	.3702	.3669	2.8611	-2.9423	-.7265	-.5393	2.2608	-.9092	1.3631	1.0724	3.7617
$\hat{\alpha}_j^0$	-4088	-2224	-1278	1022	-1839	0109	0462	3660	-3690	-0976	-0933	3150	-1119	1787	1066	4890

Note.—Each entry should be understood to have four digits to the right of a decimal point.

The values y_j and v_j are given as the final rows of Table 3.

An analysis of variance may be performed to evaluate results for this design, for which the four main effects and six first-order interactions are mutually orthogonal. Results of the analysis appear in Table 4. It is seen that all four main effects are highly significant, and that none of the interaction terms reach significance. (Higher-order interactions uniformly are insignificant, and thus are pooled with the residual mean square.) The main-effect differences may be presented conveniently in terms of "deviation contrasts" defined, for each classification, as the effect of the second class minus the mean of the effects of the two classes. For Classifications A, B, C, and D these contrasts are (A).052, (B).113, (C).150, and (D).143. The standard error of each contrast is $\sigma/4\sqrt{8}$, where σ is estimated from the residual mean square of Table 4 to be $\sqrt{.01819} = .135$. A .99 confidence interval for each contrast is available by using the t distribution, with interval boundaries given by the observed contrast plus and minus $t_{.995}(.135/4\sqrt{8})$, where $t_{.995}$ is obtained from the t distribution with 38 degrees of freedom. Confidence bounds are .032 units above and below the observed contrast. In no case does the .99 confidence interval include zero.

Interpretation of the main-class contrasts is facilitated by the way in which compensation plans were constructed. In the case of Classi-

fications A, B, and C, the benefits in the second class—weekly salary, merit incentive, and piece incentive, respectively—have approximately the same expected cost to the company, and involve no greater expected cost than the lower pay rate of Classification D. For example, the cost of fringe benefits received by salaried but not by hourly paid workers equals on the average the difference in weekly take-home pay between these two states. The contrast associated with hourly wage versus weekly salary is about 36% of that of the pay increase although the cost of these benefits to the company remains essentially the same as costs at the current, lower pay rate. Thus the average apparent value of a weekly salary is greater than its cost to the company. The apparent worth of a merit incentive based upon supervisors ratings is also greater than its expected cost to the company, being perceived as worth $.113/.143 = 79\%$ of the value of the direct-pay increase. The contrast for piece incentive shows it to be considerably greater than its cost. In fact, it even exceeds in value that of a five-cent pay increase; piece incentive has an estimated average worth $.150/.143 = 105\%$ that of a wage increase.

EXPERIMENT 2

Experiment 2 was a replication of Experiment 1, but conducted at an electrical equipment plant located on the West Coast, where working conditions differed in several respects from those at Plant 1. Two differences are particularly relevant to interpretation of findings. (a) Workers at Plant 2 were organized; no union organization had been present at Plant 1. (b) At Plant 2, piece-work incentive conditions were in effect, whereas at Plant 1, all workers were paid at hourly rates without piece-work supplements.

The stimuli presented in Experiment 2 were constructed in accordance with the same classification scheme as utilized in Experiment 1 (Table 1). However, verbal descriptions of the stimuli were modified to suit the nomenclature common to the workers at Plant 2. Further, due to a wage structure higher at Plant 2 than at Plant 1, the differential cost to the company of states within each classification was set at a level 20% greater than that in Experiment 1.

TABLE 4
ANALYSIS OF VARIANCE, EXPERIMENT 1

Source of variation	<i>df</i>	<i>MS</i>	<i>F</i>
A	1	.3487	19.17*
B	1	1.6259	89.38*
C	1	2.8990	159.37*
D	1	2.6076	143.35*
AB	1	.00003	.002
AC	1	.0050	.27
AD	1	.0444	2.44
BC	1	.0162	.89
BD	1	.0042	.23
CD	1	.0395	2.17
Residual	38	.01819	

* $p < .001$.

The Ss were 130 male employees at Plant 2. Experimental procedure was identical to that of Experiment 1. In Tables 5 and 6 appear the values p_{jk} and y_{jk} , analogous to those of Tables 2 and 3 for Experiment 1. A summary of results from analysis of variance appears in Table 7. Deviation contrasts corresponding to those given for Experiment 1 are, for the four classifications, (A).017, (B)-.110, (C).134, and (D).080. The standard error associated with each contrast in this case is $.095/4\sqrt{8}$ and a .99 confidence interval is found from the t distribution to be bounded at .022 units above and below each contrast. Consistent with results from Table 7, this confidence interval includes zero for Classification A, but not for B, C, or D.

At Plant 2, Ss do not express a significant preference for weekly wage over hourly pay (A). The large negative contrast for Classification B signifies strong rejection of a supervisory merit-rating plan as a basis for determining level of compensation. In fact, the preference for *no* such merit rating (a contrast of $-.110$) is even greater than the preference for a wage increase of six cents per hour (a contrast of $.080$). The piece-incentive feature is highly valued at Plant 2, considered on the average to be worth $.134/.080 = 168\%$ the value of a six-cent hourly wage increase.

DISCUSSION

The sharpest difference between findings from Plants 1 and 2 is that found with respect to attitudes toward a supervisory merit-rating plan. For workers at Plant 1, such a merit-rating feature is positively valued, considered worth about 80% the value of a five-cent pay raise. In Plant 2 the same feature is disliked; workers consider the *absence* of such a feature to be worth more than a six-cent pay raise. The striking difference between attitudes toward merit ratings at these two plants, whatever its cause, is of considerable interest. It is clear that an innovation in methods of compensation viewed as relatively attractive at some plants may be highly distasteful at others. (One plausible explanation of the difference in this study is the possible differential attitude toward supervisory control associated with union membership at Plant 2, contrasted with nonunion conditions at Plant 1.

TABLE 5—PROPORTIONS WITH WHICH ROW PLANS ARE PREFERRED TO COLUMN PLANS, EXPERIMENT 2

Factors ABCD	1111	1112	1121	1122	1211	1212	1221	1222	2111	2112	2121	2122	2211	2212	2221	2222
1111																
1112													.431	.492	.638	.623
1121		.538					.646	.508				.638	.538	.638		.623
1122						.715						.715				.262
1211	.415						.377				.377					
1212									.415							
1221									.492							
1222					.762						.454	.285	.654	.569		
2111			.515	.354			.523	.423				.208				
2112		.546			.746	.631										
2121	.738				.769			.654								
2122			.323							.523						
2211		.308							.415							
2212												.262			.446	.269
2221	.592		.415	.315		.577					.323					
2222										.546						

TABLE 6
ANGULAR DEVIATES AND ESTIMATED AFFECTIVE VALUES, EXPERIMENT 2

Factors ABCD	1111	1112	1121	1122	1211	1212	1221	1222	2111	2112	2121	2122	2211	2212	2221	2222
1111				-7023	1708							-4961		-0160	-1384	-1851
1112			-0761								-0921		2796	3941	0761	
1121		0761					2963			-0300			3618	2796		1708
1122	7023					4445			2963				4445		3790	2486
1211	-1708								-1708	-3618	-5144	-5681				-4961
1212				-4445			-2486		-0160	-2651	-2486				-1546	
1221			-2963			2486			-0160	-0460		-4445		1384		
1222									1546		-0921	-3131	3131			
2111				-2963	5515	1708	0160	-1546			-0460	-6237	1708			-0921
2112			0300		3618	2651	0460			0460					3618	
2121		0921			5144	2486		0921						4961		
2122	4961				5681		4445	3131	6237							
2211		-2796	-3618	-4445				-3131	-1708							-4803
2212	0160	-3941	-2796												-1082	
2221	1384	-0761		-3790		1546	-1384				-3618	-4961		1082		
2222	1851		-1708	-2486	4961					0921			4803			
β_j	-1.3671	.5976	1.1546	2.5152	-2.6627	-1.5322	-.4158	.5980	-.7170	.5648	1.3550	2.9416	-2.0501	-1.4004	-.4157	.8342
$\hat{\alpha}_j^0$	-1997	0884	1580	3076	-3453	-1971	-0464	0954	-0702	0663	1650	3652	-2287	-1846	-0727	0988

Note.—Each entry should be understood to have four digits to the right of a decimal point.

Other differences between the two plants, however, suggest that any such explanation must be considered to be highly tentative.)

The relatively stronger preference for piece-rate compensation at Plant 2 (where this feature currently is in effect) than at Plant 1 (where it is not) is consistent with additional evidence from analysis of data obtained within each plant. At Plant 2, piece-rate incentive is very strongly preferred by those workers already enjoying that benefit, but is treated with relative indifference by a subgroup of workers whose current jobs do not entitle them to piece-rate bonuses. In Plant 1, there is a tendency for workers who currently enjoy higher wage rates to appreciate the piece-incentive condition more than workers paid at a lower rate. The data suggest that a piece-incentive condition is more preferred by relatively skilled workers, and that its value is greatly enhanced after workers have had successful direct experience with the operation of such a plan.

The preference for weekly salary over hourly wage at neither plant was dramatic. Yet at both plants weekly wage was favored, significantly so at Plant 1. On the average, then, workers appear willing to forego the immediate benefits of larger take-home pay for the more remote benefits, sick leave, longer vacation, etc. For it will be recalled that the two states were designed to be of identical expected cost to the company.

The approach of this study has one char-

TABLE 7
ANALYSIS OF VARIANCE, EXPERIMENT 2

Source of variation	<i>df</i>	<i>MS</i>	<i>F</i>
A	1	.0387	4.33
B	1	1.5895	178.0*
C	1	3.8331	429.2*
D	1	.8188	91.7*
AB	1	.0011	.12
AC	1	.0026	.29
AD	1	.0077	.86
BC	1	.0032	.36
BD	1	.0361	4.04
CD	1	.0005	.06
Residual	38	.00893	

* *p* < .001.

TABLE 8
AVERAGE MONETARY VALUE, IN CENTS PER HOUR,
OF THE CONTRASTS BETWEEN PAIRS
OF COMPARABLE COMPENSATION
FEATURES

Features	Plant 1	Plant 2
A. Weekly salary versus hourly wage	\$.018	\$.013
B. Merit-rating incentive versus no merit-rating incentive	\$.040	—\$.082
C. Piece incentive versus no piece incentive	\$.052	\$.100
D. Wage increase	\$.050	\$.060

acteristic not explicitly mentioned earlier. By referring to contrasts between preference values associated with the states of each classification, and by having included a monetary pay raise as one condition to be rated, it becomes possible to compare directly the average value of each innovation with that of a pay increase. Indeed, a monetary equivalent may be attached to each preferred state of the classifications of Table 1, basing it upon the relative size of the contrast associated with that classification and the contrast associated with Classification D, the pay increase. This approach may have direct bearing upon decisions concerning whether to institute innovations in worker compensation. It would seem reasonable to approach such possible changes with a cost analysis, comparing actual cost to the company of a change in benefits with the monetary-value equivalent of that change as perceived by the workers. To the extent that workers' attitudes toward compensation are viewed as appropriate criteria for effectiveness of compensation plans, if the latter figure substantially exceeds the former, such a change might be considered to be warranted.

The present studies yield interpretation in terms of monetary-value equivalents of the compensation features as displayed in Table 8. At both plants, the average value of a change from hourly wage to weekly salary is judged equivalent to a pay increase of between one and two cents per hour. The other entries in Table 8 are subject to similar interpretation. Availability of such information provides a ready means for comparing directly on a quantitative scale the average of workers'

evaluations concerning the worth of alternative features of compensation plans, even for features which are intrinsically qualitative in nature.

REFERENCES

BOCK, R. D., & JONES, L. V. The measurement and prediction of judgmental response: Statistical methods. Chapel Hill: University of North Carolina, Psychometric Laboratory, 1963. (Mimeo)

GUILFORD, J. P. *Psychometric methods*. (2nd ed.) New York: McGraw-Hill, 1954.

McKEON, J. J. Measurement procedures based on comparative judgment. Unpublished doctoral dissertation, University of North Carolina, 1960.

THURSTONE, L. L. *The measurement of values*. Chicago: Univer. Chicago Press, 1959.

TORGERSON, W. S. *Theory and methods of scaling*. New York: Wiley, 1958.

(Received July 1, 1963)

AN INDUSTRIAL USE OF PEER RATINGS

HARRY E. ROADMAN

International Business Machines Corporation

Can peer ratings identify managers who later are promoted rapidly in a large corporation? A representative sample of 56 graduates was selected from 4-week middle manager training classes where a lengthy 13-item peer rating was administered. When the mean peer ratings of 21 managers who received 2 or more promotions after graduation were compared with the mean peer ratings of 19 managers who received no promotions, differences were significant at the .01 level for all but 3 of the rating items. The results suggest that a large body of predictive data for industry exists in the judgments contemporary managers are able to make about each other.

Industry is continuing to search for some means of identifying potentially successful top executives. Psychological tests have been reported as useful in predicting which employees will be successful junior or middle managers, especially if the tests are administered early in the individual's business career (Educational Testing Service, 1962). However, when the task is to select a few top executives from a population of apparently successful middle managers, even well-standardized intellectual test scores do not seem to have sufficient practical significance. For example, in a recent use of the Concept Mastery Test,¹ a difficult and highly verbal test of adult intellectual ability, top executives showed a significantly higher mean, 93, than middle managers, mean of 82 (the standard error of the difference between these means is 1.9). Nevertheless, the two distributions of scores overlapped extensively, with 35% of the middle managers scoring above the mean of the top executives. Thus, the inference is that one cannot predict with much confidence from a Concept Mastery Test score whether an individual middle manager will be successful as a top executive.

In contrast to the somewhat limited effort to improve systematic selection of executives (Dudek, 1963; Taylor & Nevis, 1961), larger companies are expending large amounts of money and effort in trying to train present executives and middle managers who might become executives (Pierson, 1959). At one such training program, to be described here, a tool for prediction of executive success may have been developed in the form of a peer rating involving middle managers.

¹ Kenney, R. C. Personal communication.

The extent to which peer ratings are predictive of actual executive success would appear to depend in part upon how closely the rated activities simulate real-life leadership situations. Anastasi (1961) points out in a discussion of situational techniques of assessment that, "All such tests appear to be most effective when they approximate actual 'work samples' of the criterion behavior they are designed to predict. The Leaderless Group Discussion tests (LGD), in particular, have some validity in predicting performance in jobs requiring a certain amount of verbal communication, verbal problem solving, and acceptance by peers." Of course, in the present study, leaderless problem discussion was only one of the many activities in which students got to know each other. However, in all the activities of the course, the communications, personal interactions, and decisions seemed very similar to present-day descriptions of executive job functions (Leavitt, 1958).

Whether peer ratings will predict executive success also must depend upon the reliability of these reported observations. Unfortunately, adequate reliability statistics are not available for the peer ratings reported on here. In order to encourage candid responses, the raters were not identified, so statistics of interrater agreement are not available. No subsequent peer rating has been completed with any of the student groups, so consistency of the ratings over time cannot be reported. However, some recent studies seem to bear out the statements made by authorities in psychological measurement that peer ratings are more consistent than other forms of assessment (Cronbach, 1960). For example, de Jung and Kaplan

(1962); Fiske and Cox (1960); Hollander (1957); Kaess, Witryol, and Nolan (1961); and Kubany (1957) report reliability coefficients of .60 to .70 for the consistency of peer ratings given over periods ranging up to 1 year, and reliability coefficients of .80 to .90 for interrater agreement for a single rating. As R. Bittner (1948) points out, there are several reasons why ratings made by a person's peers should reliably reflect his real competence even better than a rating by his superiors. A man's peers are usually in closer contact with what he does hour by hour and day by day, than is his superior. A man naturally tries to present his best side to his superior. Furthermore, using peers as raters makes it possible to get a large number of judgments, the average of which will be more reliable than any single measure alone.

Therefore, in this report a test is provided of the hypothesis that peer ratings can identify potentially successful executives by studying the promotion rates of various graduates of a middle manager training program in relation to the peer ratings they previously obtained during training.

METHOD

A month-long manager training course in a large corporation included a peer rating as part of a

procedure for providing career counseling to the student managers who attended the course. The peer rating was administered after each group of 16 students had been together during 2 to 3 weeks of the course. The scheduled activities during this period included lectures, discussions, problem exercises, and individual and team study projects concerning business topics such as finance, marketing, and personnel. At least 50% of the course activities were not directly structured or led by instructors, so they provided unusual opportunity for interaction between students. In addition, the student managers, most of whom did not know each other prior to the course, resided together at the school location away from their usual work places.

The peer rating procedure required each trainee to evaluate the others in his training group on 13 different items or characteristics. These characteristics are shown in Table 1. The evaluation could range from "1" (definitely possesses more of this characteristic than others in the group) to "5" (definitely possesses less of this characteristic than others in the group). The numerical ratings for each characteristic were to be accompanied by brief comments to explain the evaluation or suggest personal improvement. The rater was forced to distribute an equal number of individuals from his management training group in each of the scale values from 1 to 5. Each ratee then was awarded a mean rating for each characteristic. Each ratee was also given a rank from 1 to 16 within his class for each characteristic. None of the rating materials identified the raters in any way. Following the compilation of the results of the rating, an interview was conducted by a staff member with each student to review with him the comments of the raters about him. The peer rating

TABLE 1
PEER RATINGS FOR MANAGERS WITH DIFFERENT PROMOTIONAL RATES

Peer characteristics	Two or more promotions ^a		Number promotions ^b		<i>t</i>
	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	
Ability to think critically and analytically	4.4	2.45	11.2	4.76	5.63*
Judgment	4.8	2.80	9.8	4.29	4.30*
Originality	4.4	3.34	11.2	4.68	5.19*
Independence of thought	4.7	2.57	11.8	4.37	6.18*
Emotional maturity	6.6	5.33	8.2	3.99	1.04
Cooperation with others	10.5	5.13	8.6	4.41	1.22
Tact	8.6	5.03	7.4	4.43	.78
Aggressiveness	5.4	4.13	11.5	4.42	4.40*
Self-expression	4.8	3.46	10.8	4.38	4.71*
Leadership qualities	4.3	3.39	11.4	3.99	5.94*
Breadth of knowledge and interests	4.1	3.64	11.2	4.62	5.29*
General impression	4.4	2.82	11.1	4.34	5.70*
Capacity for advancement	4.8	3.94	10.8	4.19	4.55*

^a *N* = 19.
^b *N* = 21.
* *p* < .01.

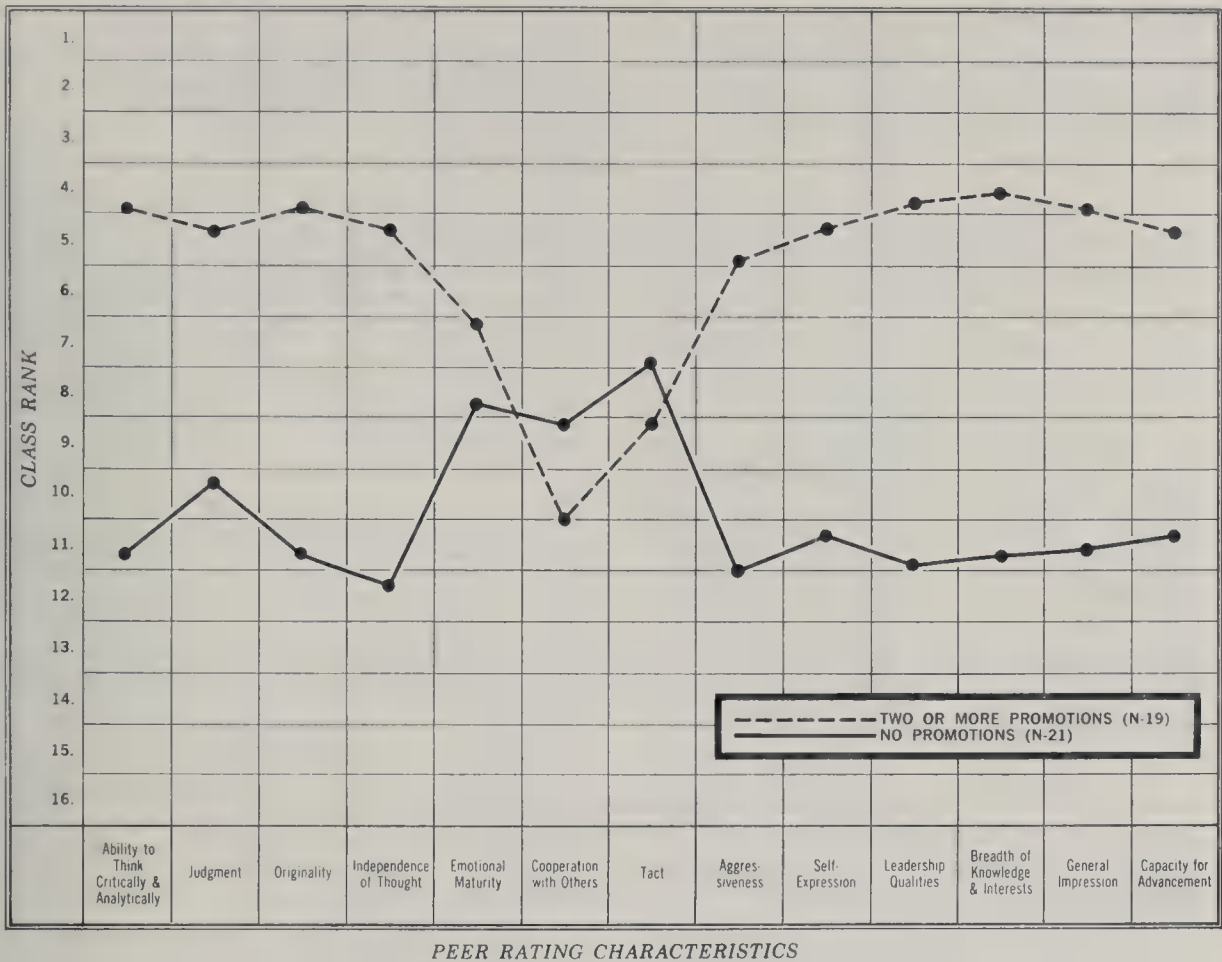


FIG. 1. Patterns of peer ratings for managers with different promotional rates.

results were then retained in confidential training folders which were not made available to the graduate's operating division.

A sample was drawn by selecting every third name from the management-school classes graduated in the 2-year period prior to 1961. This sample consisted of 56 managers who represented a variety of functional areas in the corporation such as manufacturing, engineering, finance, and marketing. Then information was screened about each of these graduates to determine if they had been given changes of assignment, which could be identified as promotions, subsequent to graduation from the management school. In view of the ambiguity often surrounding corporate organization changes, three independent opinions from experienced personnel staff members had to agree before each change of assignment was recorded as an actual promotion to greater responsibility. Three different groups of graduates were assembled according to the number of promotions received during the 2 or more years since graduation. Twenty-one managers were not promoted at all following their attendance at the management course. Sixteen managers had been promoted once during this period. Nineteen managers turned up as receiving two or more promotions in this relatively short period, with an average for this apparently successful group of three promotions.

RESULTS

Table 1 shows a comparison of mean class rank in each of the 13 peer rating characteristics for the managers who received two or more promotions and those who received no promotion. The differences between the means of these two subgroups are highly significant for all but three of the characteristics. Emotional maturity, cooperation with others, and tact are rating characteristics that do not seem to distinguish between those who are promoted rapidly and those who are not. The graph in Figure 1 shows the relative patterns of peer rating characteristics for these two different groups of managers. Such a display of the rating pattern seems useful here as a visual comparison between the peer ratings and promotional success. The graph used here is similar to the one prepared to show the peer rating for each individual graduate of the management school. Thus, it seems possible that an individual graduate's rating pattern

could be compared with the criterion pattern for those who had achieved rapid promotion in order to predict whether that individual has a high probability of such success.

DISCUSSION

The results of this study indicate that a careful and comprehensive peer rating administered in a middle manager training program can identify those who later move into senior executive positions. All of the managers in this training program had received excellent ratings in their previous assignments and had been promoted at least twice as managers in the corporation. The difficulties in assessing competence and predicting achievement among select samples of adults are well known. The peer rating as an assessment device has gained considerable attention, but the preponderance of reliability and validity studies has been with young people or in military organizations. The method and results reported here suggest that in industry also a large reservoir of predictive data exists within the observations managers make of each other's competence.

Some may be interested to know that the great majority of managers who participated in this peer rating procedure felt that it was a constructive and nonthreatening experience. Approximately 98% of the participants evaluated the peer rating and interview process as a very valuable educational experience.

Although immediate use could be made of peer ratings for some purposes of selection and placement in industrial management, additional answers are necessary. Some work is now under way in the same setting to determine whether any single peer rating characteristic, or combination of certain characteristics, may predict rapid manager promotion as well as the present pattern of rating characteristics. Also, further data are being gathered

on other measures of success, such as salary progress and supervisor's ratings, which may be used as criteria for peer rating studies. In addition, plans are being drawn to study the relationship these peer ratings have to other selection factors such as early performance appraisals, training staff ratings, and objective test scores. Finally, some attention must be directed to the question of the willingness of participants to accurately rate each other if they know full use will be made of peer ratings for administrative as well as individual development purposes.

REFERENCES

- ANASTASI, A. *Psychological testing*. New York: Macmillan, 1961.
- BITTNER, R. Developing an employee merit rating procedure. *Personnel Psychol.*, 1948, **1**, 403-432.
- CRONBACH, L. J. *Essentials of psychological testing*. (2nd ed.) New York: Harper, 1960.
- DE JUNG, J. E., & KAPLAN, H. Some differential effects of race of rater and ratee on early peer ratings of combat aptitude. *J. appl. Psychol.*, 1962, **46**, 370-374.
- DUDEK, E. E. Personnel selection. *Annu. Rev. Psychol.*, 1963, **14**, 261-279.
- EDUCATIONAL TESTING SERVICE, The conference on the executive study. *Identifying management talent*. Princeton: ETS, 1962.
- FISKE, D. W., & COX, J. A., JR. The consistency of ratings by peers. *J. appl. Psychol.*, 1960, **44**, 11-17.
- HOLLANDER, E. P. The reliability of peer nominations under various conditions of administration. *J. appl. Psychol.*, 1957, **41**, 85-90.
- KAESS, W. A., WITRYOL, S. L., & NOLAN, R. E. Reliability, sex differences, and validity in the LGD technique. *J. appl. Psychol.*, 1961, **45**, 345-350.
- KUBANY, A. J. Use of sociometric peer nominations in medical education research. *J. appl. Psychol.*, 1957, **41**, 389-394.
- LEAVITT, H. J. *Managerial psychology*. Chicago: Univer. Chicago Press, 1958.
- PIERSON, F. C. *The education of American businessmen*. New York: McGraw-Hill, 1959.
- TAYLOR, E. K., & NEVIS, E. C. Personnel selection. *Annu. Rev. Psychol.*, 1961, **12**, 389-409.

(Early publication received August 27, 1963)

SELECTING COMPETENT RATERS¹

LLEWELLYN WILEY AND WILLIAM S. JENKINS

Aerospace Medical Division, Lackland Air Force Base, Texas

Navigator students ($N = 109$) estimated qualifications needed to perform Air Force tasks using an experimentally standardized list and sets of 5 rating scales. Rerating a month later, they were scored by correlating their ratings with a key of pooled estimates. Key agreement significantly predicted key agreement, suggesting a technique for selecting task analysts and job evaluators.

Personnel psychologists find it necessary to use human judgments to obtain estimates of the qualifications required in the successful performance of jobs. The reliability of the estimates of a few people becomes very important for those jobs which have extremely few experienced personnel, as, for example, in estimating the qualifications for an astronaut's duties. For more populated occupations one must select among many experts to economize time and data. In our present research we are attempting to identify reliable judges and raters by developing a standard yardstick for appraising their ability to estimate job requirements.

This instrument lists 45 Air Force tasks and the names of 15 specialties from which they were taken, all reasonably familiar to the student officers who have served as raters.² There are three sets of rating scales which can be applied to the lists of tasks and specialties.³ In each set there are five 5-point scales. A rater can rate all 60 items (45 tasks and 15 specialty names) of the list in an hour's time, producing 300 estimates of the qualifications required to perform representative jobs. Combinations of parallel lists and scales permit study of rating tendencies and individual rater

reliabilities. In successive studies we found consistencies after 4 hours, 4 weeks, and 10 months, tending to repeat individualistic high, low, or extreme ratings (Wiley, 1963; Wiley & Jenkins, 1963). The same instruments were turned to a different purpose in the present study: to determine if it is possible to identify those raters who continue to represent the group consensus after the lapse of a month's time. The study was feasible because the rating instrument has been designed in multiple forms.

The concept of rater agreement with a consensus was applied by Dubin, Burke, & Katz (1954) to ratings made by noncommissioned officers on overall qualities of enlisted men. They obtained considerable reliability among raters on the evaluation of people. Mullins and Force (1962) also studied rater agreement in rating personnel, concluding that the ability to agree with the consensus is generalizable across rating dimensions. The raters in the present study were not evaluating individuals, but estimating the qualifications demanded of abstract workers in the performance of tasks and specialties. The study therefore supplements rather than duplicates the small body of existing knowledge concerning rater agreement with a consensus.

The present study follows these steps:

1. The means of a pool of ratings made on each item on each rating scale constituted a scoring key for that scale. There were two sets of keys.
2. A rater was scored by correlating the ratings he gave to each item on each scale with the values in the scoring key. With five scales he received five correlation coefficients, or scores. There were two rating sessions,

¹ The research reported in this paper was sponsored by the 6570th Personnel Research Laboratory, Aerospace Medical Division (AFSC), Lackland Air Force Base, Texas under AFSC Project 7734, Task 773404.

² There are currently three forms of the task list, A, B, and C. Only C is available for release, and only C was used in this study. A copy of Form C may be obtained for research purposes by requesting PL 9144 from the 6570th Personnel Research Laboratory.

³ There are three forms of groups of five scales, Forms U, V, and W, which may be obtained in one copy each for research purposes by requesting PL 9160, PL 9161, and PL 9162 from the 6570th Personnel Research Laboratory. Only Forms U and W were used in this study.

yielding two sets of scores, one for each set of keys.

3. Correlation coefficients were converted to Fisher's z scores so that they could be handled like ordinary numbers. A rater now received two mean scores, representing his average agreement with the group consensus on two occasions a month apart.

4. The first score was used to predict the second by ordinary correlation. Determining the extent of this prediction was the major objective of the study.

METHOD

Subjects

Small groups of navigator students were tested as raters at their training base over a 3-month period. Each group was tested a second time after about a month, on one occasion with scale Form U, on the other occasion with scale Form W. On both testings they were required to rate the 45 tasks and 15 specialties of Task List C on all five scales. Due to omissions and to the "90% of subjects completing" rule applied in the testing, data were incomplete for many raters. A few nearly complete answer sheets were retained by inserting the scale midpoint, 3, for missing data. Usable data were obtained for 109 raters (58 student officers and 51 aviation cadets). In the order in which they applied the scales, 67 raters employed Form U initially, following a month later with Form W, while 42 raters started with Form W and followed with Form U. The 42 raters who started with scale Form W did not enter into the pool which produced the scoring keys, but the 67 who started with Form U did. The remaining contributors to the consensus pool were obtained from those raters for whom data were incomplete on one or the other session. Thus, the raters who augmented the consensus data for the Form U key were necessarily different individuals from those who augmented the Form W pool.

Procedure

A rater received a score obtained by correlating the ratings he assigned to each of 60 items on each scale with the array of means of a group of raters applying the same scale to the same items, i.e., the key. This correlation was converted to Fisher's z score. A rater thus received five z scores on each rating session, one for each scale, and each reflecting a correlation coefficient of 60 observations. The central problem was to establish the extent to which the mean of the five z scores predicts the mean of the corresponding scores, indicating the stability of the ordering of the 109 raters in respect to their agreement with the group consensus.

RESULTS AND DISCUSSION

Correlations discussed hereafter were obtained from prediction of Fisher z scores by Fisher z scores. The problem (Will agreement predict agreement?) is attacked by computing the mean of z scores for the form used initially and correlating these with the corresponding mean z scores on the second testing. The first mean z score predicted the second mean z score with a r of .44 ($r = .25$ significant at the .01 level). This answers the basic question of the study: there is significant prediction of rater agreement with a consensus.

We were interested in learning whether the presence of repeated scales could be made to account for all the predictability of consensus agreement. The resourcefulness and precision scales had been repeated, three other scales in Form U and two others in Form W had not (one scale in Form W was dropped from the study). By combining first and second testing data, all data (correlations) for Form U could be collected and studied; likewise, the data for Form W could be assembled. The mean of a rater's correlations with the resourcefulness and precision keys (repeated scales) could be correlated with the mean of his correlations with the keys of the nonrepeated scales. This procedure yields something analogous to an uncorrected reliability of agreement scores; it indicates the similarity of scoring a rater on nonrepeated scale agreement to scoring him on repeated scale agreement. For Form U the resulting r was .61; for Form W it was .70 ($N = 109$; Fisher z transformations used). The finding suggests that there is considerable consensus agreement across time which is independent of the repetition of scales.

We next asked the question: Does the rater who agrees with the key also tend to be the one who agrees with himself? Raters could be correlated with themselves across time on the repeated scales, resourcefulness and precision (two r 's of 60 observations which are independent of a key). The mean of these two retest reliabilities did not differentiate the raters as well as a single r obtained by treating both scales as though they were only one, with 120 observations. A corresponding r (120 observations) was computed for each rater's agreement with the combined resourcefulness

and precision keys. After transformation to Fisher z scores, the key-agreement score was used to predict the self-reliability score. This r was .64 (109 observations). It appears that raters who tend to agree with the consensus also tend toward retest self-agreement.

CONCLUSIONS

The findings are evidence that a rater's correlation with a consensus key can be used to predict both his self-consistency and his agreement with consensus keys on later testing. The prediction possible with an r of .44 is not very great, but in this instance it was found that useful selections could have been made among the 109 raters sampled; 14 of the 27 raters composing the upper quartile on the first testing reappeared in the upper quartile on the second. These findings are limited to

ratings on the qualifications estimated to be required of abstract persons who are to perform jobs. It would be interesting to learn whether parallel consistencies over time can be demonstrated in raters who rate real people.

REFERENCES

- DUBIN, S. S., BURKE, L. K., & KATZ, A. *Characteristics of raters whose ratings agree with consensus ratings*. Washington, USAF PRS tech. res. Note, 1954, No. 35.
- MULLINS, C. J., & FORCE, R. C. Rater accuracy as a generalized ability. *J. appl. Psychol.*, 1962, **46**, 191-193.
- WILEY, L. Influence of time on rater bias measurements when estimating task requirements. *J. industr. Psychol.*, 1963, **1**, (2), 39-43.
- WILEY, L., & JENKINS, W. S. Method for measuring bias in raters who estimate job qualifications. *J. industr. Psychol.*, 1963, **1**, (1), 16-22.

(Received June 18, 1963)

PREDICTION OF PERFORMANCE IN MEDICAL SCHOOL FROM THE CALIFORNIA PSYCHOLOGICAL INVENTORY¹

HARRISON G. GOUGH AND WALLACE B. HALL

University of California, Berkeley

Prediction of performance in medical training is a difficult task, and few approaches to date have shown promise. In the nonintellective realm, the California Psychological Inventory (CPI) has given positive results in other settings and was therefore deemed worthy of tryout in this one. A CPI regression equation was derived which had a predictive validity of $+0.66$ in an initial sample of 34 students, and of $+0.46$ in a cross-validating sample of 63. The psychological dimension defined by the equation was judged to be more reflective of a characterological syndrome stressing unselfishness and consideration for others than of need:achievement, compensatory striving, or other factors typically invoked to account for scholastic attainment.

Forecasting of performance in medical training is a problem of obvious importance to applied psychology. Although it is performance in medical practice which is the fundamental outcome, achievement during training may nonetheless be viewed as a proximate criterion; the training period must be successfully completed before the student can go on into internship and practice, and indices of achievement during the academic period appear to function fairly well as predictors of performance during the internship (cf. Richards, Taylor, & Price, 1962).

Attempts to predict performance during medical training have typically drawn on three sources of evidence: (a) intellectual aptitude, primarily as measured by the Medical College Admission Test (MCAT) (see Wantman, 1953; Wesman, 1959); (b) premedical scholastic achievement; and (c)

ratings made by interviewers at the time of application.

The available evidence (see Gough, Hall, & Harris, 1963) suggests that these sources of information do not predict achievement in medical training very well. With respect to the MCAT, the reviews by Wantman (1953) and Wesman (1959) both stressed the lack of predictive utility (or, more precisely, the lack of any evidence concerning the predictive utility) of the test. Research inquiries such as those of Richards and Taylor (1961), Kelly (1957b), and Hammond and Kern (1959) document these negative judgments.

Kelly (1957b), for example, factored 32 criterion variables for a graduating class of 112 medical school students at Michigan, identifying five major dimensions. Median correlations with each factor for the four subtests of the MCAT were: verbal, $+0.07$; quantitative, $+0.05$; modern society, $+0.01$; and mastery of premedical sciences, -0.02 . Garfield and Wolpin (1961) found the MCAT unable to distinguish between 51 dropouts and 626 graduates in eight consecutive classes at Nebraska. Schwartzman, Hunter, and Lohrenz (1962, p. 751) likewise concluded that "the relation between the MCAT and medical school performance is not strong." Perhaps the most forceful statement is that of Wesman (1959, p. 937) who wrote, "The over-all picture of validity provokes one to question whether the individual medical

¹ This study is one of a series of investigations on interpersonal effectiveness and creative achievement being conducted at the Institute of Personality Assessment and Research, University of California, Berkeley. Work on the study was supported directly by Grant No. 5746, National Institute of Mental Health, and by research funds made available by John B. deC. Saunders, MD, Provost of the University of California, San Francisco, School of Medicine, and indirectly by a Ford Foundation grant for basic research in the behavioral sciences to the senior author. We also wish to acknowledge the advice and counsel of Robert H. Credé, MD, Professor of Medicine and Associate Dean of the University of California School of Medicine, San Francisco.

schools are (or should be) satisfied with the program."

Premedical grades seem to be somewhat more useful as predictors of performance in medical school than the MCAT. Gottheil and Miller (1957) suggest a mean validity coefficient of $+0.50$; however, their survey included studies of achievement during the first 2 years of training, when basic sciences predominate; a lower value would be expected for achievement during the last 2 years when clinical practice receives greater emphasis. Against overall grades in medical school, Hammond and Kern (1959) report a coefficient of $+0.29$ for premedical grade-point average (GPA), and in Kelly's (1957b) study the median coefficient for premedical GPA against the five factors was $+0.18$.

Psychometric evidence on the predictive value of the admissions interview is less readily available. In a general consideration of this interview, Kelly (1957a) concluded that it was almost entirely lacking in validity. In his study at Michigan, Kelly (1957b) found that the median correlation between interviewers and the five factored criteria was only $+0.05$. Gough, Hall, and Harris (1963) also found a median coefficient of $+0.05$ between interviewers' ratings and overall GPA in medical school in a study of 14 medical school classes at the University of California, San Francisco.

Definitive conclusions concerning a literature as vast as that dealing with forecasting performance in medical training may be impossible. Certain trends, however, are clear enough and serve to indicate that the MCAT, the admissions interview, and premedical grades have much lower predictive value than most psychologists would expect. Indeed, the MCAT and the interview may be hard put to exceed a mean value of $+0.20$ if overall grades and faculty evaluations at the time of graduation are used as criteria.

This state of affairs has led various investigators to search for other sources of prediction, particularly among "non-intellectual" variables (see Funkenstein, 1957). Eron (1954), for example, administered the Rorschach test to 35 medical school students and 35 divinity students. Although judges were able to differentiate between the two

samples with low reliability, they were not able to distinguish the protocols of 10 high-ranking medical students from those of 10 low-ranking students. Eron (p. 39) concluded that the use of the Rorschach in the selection of medical students or in the prediction of their success is not justified.

Schofield (1953) applied the MMPI to a sample of 83 freshman medical students. On the basis of grades in the junior year, two subsamples (matched on the A.C.E. test of scholastic aptitude) were constituted, 11 students in the top quarter on grades and 11 from the bottom quarter. Four scales revealed significant differences (*L*, *Hy*, *Pd*, and *Sc*), with the high-ranking students having lower means in each instance. Unfortunately, correlations of MMPI scales with grades for the full sample of 83 were not presented.

Two studies, however, have offered MMPI correlational data for total classes. Glaser (1951), in a sample of 150 medical students, found no significant correlations between MMPI scales and first-year grades. Glaser also considered the "sign" approach, but failed to identify any MMPI signs having significant relationships to scholastic achievement. Knehr and Kohl (1959) studied 249 students in three graduating classes. MMPI scales were correlated with cumulative class standings, but no significant values were obtained. They then identified 63 students whose MMPI profiles were indicative of psychiatric instability, but found that only 18 of these students had later received psychiatric treatment while in school. Nineteen other students, whose MMPI profiles did not indicate such problems, had also gone on to seek psychiatric consultation.

The Strong Vocational Interest Blank (SVIB) would seem to be a promising candidate as a forecaster of performance in medical training, but in Stuit's (1941) study of 131 students the physician's key correlated only $+0.16$ with first-year grades. In the study by Holt and Luborsky (1958) on psychiatric residents, the SVIB was able to differentiate between accepted and rejected residents (p. 41), but did less well in forecasting differential attainment among 64 subjects in the final predictive study. In this study neither the clinical-psychologist key nor the psychia-

trist key correlated significantly with any major criterion. The 1948 psychologist key correlated $+.21$ and $+.22$ against peers' and supervisors' evaluations of competence in diagnosis and therapy, and the lawyer key correlated $+.22$ and $+.23$ with two criteria of psychotherapeutic competence.

In light of the discouraging history sketched above the reader may wonder why anyone would wish to continue the search for predictors—intellectual or nonintellectual—of performance in medical training. Other tests and procedures, nonetheless, do await tryout as predictors and it is always possible that new and promising leads will be discovered.

One such possibility is the California Psychological Inventory (CPI) (Gough, 1957), which was originally constructed in the hope of providing measurement of those interpersonal factors of character and temperament which are involved in everyday social living and constructive achievement; i.e., the CPI is addressed to variations within the positive sphere of ego functioning (which, it should be stressed, is not merely equivalent to an absence of psychopathology and malfunctioning).

Recent literature has begun to show that the CPI does have predictive value in various educational settings. Holland (1959) found, in his study of National Merit Scholarship Corporation finalists, that the CPI yielded predictive validities significantly superior to those derived from aptitude test scores. Maxwell (1960) found CPI patterns to forecast graduation versus dropping out from college, and Keimowitz and Ansbacher (1960) found significant correlations between the CPI and differential achievement in mathematics.

There is also some evidence on differential performance within specific occupations. Barron (1963) found that 10 of the 18 scales of the CPI differentiated between creative writers and controls, and MacKinnon (1962), in his study of architects, observed significant differences between more- and less-creative men on 9 of the scales from the inventory.

Because of the more or less "open" state of the prediction problem in medical education, and because of the promise shown by the CPI in other settings, it was decided to undertake an analysis of performance in

medical school, using the CPI as a predictive instrument.

PROCEDURE

In the fall of 1954, 70 male applicants to the University of California School of Medicine in San Francisco were studied in an assessment program at the Institute of Personality Assessment and Research in Berkeley. An additional group of 30 male applicants participated in a testing program, but did not undergo the full assessment program. Thirty-four of these 100 applicants later entered and graduated from the training program.

For these 34 men, MCAT and CPI scores and pre-medical grade-point averages at the time of application were assembled. During the course of training, grade averages were calculated each year and for the overall 4-year period. In addition, in the junior (1958) and senior (1959) years, faculty ratings on performance and potentiality were obtained for all students in the class and factor scores derived; a single, composite faculty rating summed over all individual factors was used as the criterion.

The initial sample therefore consisted of 34 students for whom three sources of prediction (CPI, MCAT, and premedical GPA) and six criteria (GPA by year, overall GPA, and faculty rating) were available. The time span between testing and the most remote criterion (evaluation at the time of graduation) was from the fall of 1954 to June of 1959.

The cross-validating sample consisted of 63 males from the 1956 graduating class of the University of Colorado School of Medicine.² These students were tested with the CPI a year or more prior to graduation, and the 4-year GPA in medical school was taken as the criterion.

RESULTS

Means and standard deviations on the CPI and MCAT for the 34 students in the initial sample are shown in Table 1, along with corresponding data for the 66 applicants who were not admitted. Only one variable, the socialization scale of the CPI, differentiates the samples at the .01 level; a second variable, the quantitative scale of the MCAT, differentiates the two samples at the .05 level. Although the MCAT scores for these 100 applicants were available to the Admissions Committee, it would appear that other sources of information (viz., premedical grade averages and ratings by the Admissions Com-

² These data were made available through the courtesy of Kenneth Hammond, Director of the University of Colorado Behavior Research Laboratory.

TABLE 1

COMPARISON OF APPLICANTS WHO WERE ADMITTED TO MEDICAL SCHOOL AND LATER GRADUATED
WITH APPLICANTS WHO WERE NOT ADMITTED

Variable	Graduates ^a		Not admitted ^b		diff	t
	M	SD	M	SD		
CPI						
Do	32.29	5.49	30.71	5.06	1.58	1.44
Cs	24.09	2.62	23.41	3.19	.68	1.07
Sy	29.09	3.11	28.23	4.01	.86	1.09
Sp	39.24	5.32	39.41	5.30	-.17	.16
Sa	23.76	2.51	23.83	3.25	-.07	.11
Wb	40.09	2.67	39.92	2.82	.17	.28
Re	34.29	4.09	33.98	3.57	.31	.39
So	41.26	4.21	38.47	5.03	2.79	2.78**
Sc	32.21	5.98	32.35	6.58	-.14	.11
To	27.82	3.79	27.00	3.13	1.82	1.16
Gi	22.32	5.69	20.50	6.28	1.82	1.42
Cm	26.18	1.42	26.30	1.49	-.12	.41
Ac	31.74	3.63	31.23	2.91	.51	.76
Ai	23.59	3.74	23.91	3.23	-.32	.45
Ie	43.74	3.08	44.17	2.90	-.43	.69
Py	13.38	2.64	13.79	2.12	-.41	.83
Fx	11.71	3.91	11.56	3.67	.15	.18
Fe	16.29	3.78	16.45	3.37	-.16	.22
MCAT						
Verbal	553.82	72.40	536.82	81.47	17.00	1.03
Quantitative	549.71	54.13	512.42	74.26	37.29	2.59*
Modern society	551.18	76.58	535.46	80.14	15.72	.94
Science achievement	560.29	76.06	530.61	70.40	29.68	1.94

^a N = 34.^b N = 66.* $p < .05$.** $p < .01$.

mittee interviewers) were weighted more heavily in determining admission.

For the 34 admitted candidates, the means on the MCAT are up about half a standard deviation over the estimated national average for all applicants to medical schools; the standard deviations reveal a reduction of from 23 to 46% from the presumed national norm of 100.00. Some restriction of range on the MCAT has therefore occurred for the 34 students included in the initial sample. The reader, however, should not lose sight of the fact that the processes of selection (in particular the Admissions Committee interviewers' ratings) will also lead to a restriction of range on the variables scaled in the personality inventory. For example, the standard deviation of 4.21 on the socialization scale for the 34 selected students is about 20% less than the value cited in the *CPI Manual* for

college males, and the mean score of 41.26 is approximately three-fourths of a standard deviation above the national average for college men.

The CPI figures in Table 1 are in raw score form. If expressed as standard scores, the highest values (above 60 on the individual profile sheet) for the 34 admitted students would be on the scales for Achievement via Independence (Ai), Capacity for Status (Cs), Dominance (Do), Self-Acceptance (Sa), and Tolerance (To). The lowest point would be on Femininity (Fe), where the standard score would be 50 (the mean for men-in-general), and the next lowest point would be on Self-Control (Sc) where the standard score would be about 51.

Correlations of the predictive variables with the six criteria are shown in Table 2. Of the 108 correlations computed for the CPI (18

scales times six criteria), 11 are significant at or beyond the .05 level of probability. Of greatest interest are the relationships with overall GPA and the faculty ratings. The Sociability (Sy), Tolerance (To), and Intellectual Efficiency (Ie) scales reveal significant positive correlations with overall GPA, and the Sociability (Sy) scale has a correlation of +.48 with the ratings. These individual coefficients are provocative, and offer preliminary leads for interpretation, but would need cross-validation before being seriously considered. Also, they carry little or no information concerning the combination or pattern of CPI variables which would best predict the various criteria.

None of the 24 correlations for the MCAT (six criteria times four scales) attains the .05 probability level, although the coefficient of

+.28 between the science subtest and first-year grades is not too far below this point. However, even the science subtest, by the time the faculty ratings are reached, is not functioning very well as a predictor ($r = -.14$). Premedical scholastic achievement, as calibrated in three different ways, fares no better. None of the 18 correlations is significant, and the three coefficients with the faculty ratings ($-.11$ for overall premedical GPA, $+.08$ for premedical GPA in sciences, and $-.18$ for premedical GPA during the last two terms of study) offer little encouragement for multiple-regression analyses.

The variations in Table 2 from column to column may reflect, to some extent, error and sampling variation within the predictors, but there are also some meaningful trends discernible among the criteria. Table 3 presents

TABLE 2
CORRELATIONS OF PREDICTORS WITH GRADES IN THE FOUR YEARS OF MEDICAL TRAINING AND WITH FACULTY RATINGS, FOR A SAMPLE OF 34 STUDENTS

Variable	1st year GPA	2nd year GPA	3rd year GPA	4th year GPA	Overall GPA	Faculty ratings
CPI						
Do	.15	.34*	.24	.25	.30	.31
Cs	.16	.01	.10	-.20	.05	-.06
Sy	.13	.16	.41*	.37*	.35*	.48**
Sp	.30	.02	.28	.01	.23	.11
Sa	.12	.08	.23	.34*	.26	.29
Wb	.10	.10	.28	.24	.23	.23
Re	.00	.28	.18	.12	.16	.10
So	.09	.03	-.03	-.16	-.04	-.08
Sc	.05	.21	.22	.15	.18	.10
To	.09	.33	.42*	.23	.34*	.29
Gi	.07	-.04	.14	.00	.06	.04
Cm	-.13	-.01	.01	.02	-.06	.32
Ac	-.06	.03	.19	.23	.11	.22
Ai	-.08	.11	.12	.00	.06	.07
Ie	.25	.28	.38*	.29	.40*	.22
Py	.34*	.23	.30	.06	.32	.10
Fx	.04	-.07	.00	-.20	-.03	-.14
Fe	.04	.09	.05	-.08	.03	.15
MCAT						
Verbal	.00	.14	-.04	-.01	.03	-.18
Quantitative	-.01	-.23	-.03	-.05	-.08	.00
Modern society	.01	.21	.06	.13	.12	.01
Science	.28	.24	.04	.04	.18	-.14
Premedical GPA						
Overall	.00	.08	-.06	.08	-.02	-.11
Science	.16	.26	.14	.10	.21	.08
Last two terms	.02	-.02	-.07	-.13	-.06	-.18

* $p \leq .05$.
** $p \leq .01$.

the intercorrelations among the criteria. Note that grades in the first 2 (preclinical) years correlate only +.19 and +.44 with grades in the fourth year, and only +.12 and +.44 with the faculty ratings. Achievement in the first 2 years of medical school is not a very good predictor of faculty ratings at graduation, in this sample.

The practice in many studies of using first-year grades in medicine as the principal or only criterion seems questionable, if the data in Table 3 can be considered as generalizable. Grades in the third and fourth years, on the other hand, when clinical practice is stressed, show higher correlations (+.73 and +.68) with faculty evaluations on the factorial criterion.

The next step in our analysis was the derivation of various multiple-regression equations, seeking optimum combinations of scales from the CPI and the other predictors to forecast performance at each stage of medical training. Because of an interest in a direct comparison between the CPI and MCAT, it was decided to restrict the number of scales in any CPI equation to four, equating this to the four scales in the MCAT.

Six regression equations for the CPI were developed in this way, one for each of the six criteria, and likewise six equations were developed for the MCAT. Each CPI equation, as mentioned, included four scales. Correlations between each equation and its own criterion were as follows: first-year grades, $r = +.50$; second-year grades, $r = +.50$; third-year grades, $r = +.57$; fourth-year grades, $r = +.55$; overall GPA, $r = +.49$; and faculty ratings, $r = +.66$. There is a slight increase in the multiple correlations in the clinical years over the preclinical years, and the highest single value is observed for the faculty ratings. In other words, as the criterion moves closer to the circumstances of professional practice the validity of the CPI seems to improve.

A somewhat different set of findings was observed for the MCAT. Although all four scales were considered at each stage, in only one (second-year grades) did any combination of scales exceed the accuracy of prediction which was given simply by taking the scale with the highest correlation. In addition to

TABLE 3
INTERCORRELATIONS AMONG THE CRITERIA OF GRADES
IN MEDICAL TRAINING AND FACULTY RATINGS, IN
A SAMPLE OF 34 STUDENTS

	2	3	4	5	6
1. First-year grades	.55	.43	.19	.74	.12
2. Second-year grades	—	.60	.44	.80	.44
3. Third-year grades		—	.64	.84	.73
4. Fourth-year grades			—	.71	.68
5. Overall grades				—	.62
6. Faculty ratings					—

Note.—For $N = 34$, $r_{p.05} = .34$; $r_{p.01} = .45$.

low individual validity, the MCAT scales have rather high intercorrelations with each other. Keeping in mind that only for second-year grades is more than one MCAT scale utilized in the forecast, the correlations between MCAT and the criteria are these: first-year grades, $r = +.28$; second-year grades, $r = +.27$; third-year grades, $r = +.06$; fourth-year grades, $r = +.13$; overall GPA, $r = +.18$; and faculty ratings, $r = -.18$.

This last value, $-.18$, merits comment as it indicates that for this sample of 34 the students rated highest by the faculty tended to have lower verbal scores than students rated as less promising. It might also be remarked that whereas the CPI improves as a predictor of performance in training as one approaches the time of graduation, the MCAT declines in validity.

All of the above, albeit interesting, would hardly be worth attention unless cross-validating evidence could be presented. Fortunately, the sample of 63 graduating seniors from the University of Colorado School of Medicine is available. In line with the belief that the faculty ratings offer the most important and significant criterion among the six available, it was decided to consider only the equation developed for its prediction. That equation is offered below:

Medical promise = + .794 Sy + .602 To
+ 1.144 Cm - .696 Cs [1]

This equation was applied to the raw scores of the CPIs for the 63 Colorado students, and a predicted score calculated for each man. This distribution of 63 predicted scores had

a mean of 50.06 and a standard deviation of 4.97.

The CPI "medical promise" scores were then correlated with the 4-year GPA, yielding a coefficient of $+ .46$, significant well beyond the .01 level of probability. For these same 63 students, the correlations between 4-year GPA and the MCAT were: verbal, $r = + .12$; quantitative, $r = + .03$; modern society, $r = + .10$; and science, $r = + .23$; and for 4-year GPA versus premedical GPA the coefficient was $+ .18$. Thus, the cross-validating coefficient of $+ .46$ for the CPI equation is not only significant statistically, but is considerably higher than what was observed in the same sample for the MCAT or premedical grades.

PSYCHOLOGICAL MEANING

Having demonstrated, at least to some extent, the predictive validity of the equation from the CPI the next question becomes one of its psychological meaning. From its mode of derivation we know that it identifies, within a margin of error, students who will do well in medical training. Its components include a positive weighting of Sociability, Tolerance, and Communality, and a negative weighting of the status scale. This information, although relevant, is not sufficient to permit an interpretation of the "psychology" of the equation. We have therefore sought new and additional information.

This new information is of the kind offered in the *CPI Manual* (Gough, 1957) for scales of the inventory, but here offered for a regression equation composed of CPI scales. It consists of descriptive data about persons with higher and lower scores on the equation, and is intended to answer the question "What kind of a person is it, in everyday description and language, who is identified as a high-scorer by this equation?" To say that he is a person who ought to do well in medicine, and who has higher standing on Sy, To, and Cm, and lower on Cs is only a partial answer to the question; a more comprehensive answer must include study of what subjects are doing in life as well as what they have done on the test.

The sample on which this personological study was conducted consisted of 41 members

of a University of California fraternity.³ The sample may be deemed appropriate to the aims of the study, as the students were mostly juniors and seniors in college, of about the same age and educational status as the modal applicant to medical school. The question being considered, if a reminder be permitted, is what psychological and behavioral characteristics can we expect of an intelligent young man, junior or senior year in college, if he attains a high score on the equation? Our high-scoring subject may or may not be an applicant for medical school, but whether he is or not we wish to know the personological implications of his standing on the distribution of scores defined by the equation.

Each boy took a series of tests and questionnaires, including the CPI, and the regression score for medical promise was calculated. For this sample the mean was 49.27, *SD* 5.37. Then, at a later time, each subject was asked to complete an Adjective Check List (ACL) (Gough, 1960) on five other subjects. For each ACL, the observer was asked to check those words which he considered to be salient and descriptive of the subject, and then to double-check the 10 most highly descriptive terms. Each of the 41 boys was in this way described by five of his peers, and for each of the 300 adjectives in the ACL a "score" was tallied by counting the number of checks it had received. For example, if the five ACLs turned in on "Anderson" by his peers showed three single checks on the word "alert" and one double-check, Anderson's score on alert was tallied as five. If "Bates" was single-checked twice on "blustery" and double-checked three times, his score on blustery would be entered as $8(2 \times 1 + 3 \times 2)$.

These scores, 300 for each man, can be viewed as social-psychological descriptions of the everyday social and interactional behavior of each subject. The procedure of analysis was simply that of correlating the CPI predicted score on "medical promise" with these 300 adjectival descriptions. Please note that the ACL descriptions are *not self-reports*, but observations made about each subject by his peers.

³ We wish to thank the members of the Phi Sigma Kappa fraternity at the University of California for their participation in this phase of the study.

In a sample of 41 cases, correlations of $\pm .30$ or more are significant at or beyond the .05 level. By chance, one would expect 15 of the 300 correlations with the ACL to reach or exceed this level. In fact, 23 equaled or surpassed this value. The adjective having the highest positive correlation ($+.39$) was *unselfish*, and the adjective with the largest negative value ($-.38$) was *thankless*. The next 5 positive correlations, in order, were *considerate* ($r = +.35$), *informal* ($r = +.35$), *forgiving* ($r = +.34$), *reasonable* ($r = +.34$), and *self-confident* ($r = +.34$). On the negative side, adjectives were *cold* ($r = -.37$), *cool* ($r = -.37$), *prejudiced* ($r = -.37$), *fault-finding* ($r = -.35$), and *restless* ($r = -.35$).

The "syndrome," therefore, defined by high scores on the CPI equation is one in which traits such as unselfishness, considerateness, informality, reasonableness, self-confidence, and forgivingness are prominent. The low-scorer on the equation is a young man who tends to be thankless, cold, cool, prejudiced, fault-finding, and restless. The CPI equation identifies, that is to say, a kind of person who is responsive to others, charitable, and yet sure of himself. Note that words such as ambitious, enterprising, intelligent, persevering, etc., which are on the ACL and which might be thought to pertain to performance in medical training, are not implied by the equation. The equation, i.e., does not stress a need: achievement pattern, an intellectuality pattern, a creativity pattern, or a personal aggrandizement pattern. What it does betoken is a pattern of personal resourcefulness coupled with sensitivity to the needs and demands of others. And what it "screens out" (low scorers on the equation) is a kind of petulance, self-centeredness, and intolerance which most would see as undesirable characteristics in a prospective medical practitioner.

INTERPRETATION AND DISCUSSION

Although the prediction of performance in medical training is a matter of obvious importance, studies of the problem have typically yielded insignificant and inconsequential results. Even the procedures used to evaluate applicants for admission—aptitude test scores,

premedical scholastic achievement, and interviewers' ratings—have not fared very well.

Experimentation with new and different techniques is therefore in order, and should be welcomed. The so-called nonintellective domain is one in which much current effort is being expended, but even here the yield has not been impressive. The CPI, which has shown considerable promise in forecasting achievement in other settings and which seeks to measure dimensions of ego effectiveness, would seem relevant to the prediction of performance in medical training and hence worthy of trial.

The predictive results in the two samples studied were positive, and the magnitudes of the relationships high enough to give promise of practical utility. The CPI does, therefore, appear to tap personological dispositions significantly involved in successful performance during the course of medical training. It must, of course, be recognized that the samples giving rise to these impressions are small and that the circumstances of testing (special research surveys) were such as to minimize distortions and other difficulties which would be encountered in an operational program.

Perhaps of greatest immediate interest is the psychological nature of the personality syndrome defined by the CPI equation. This syndrome appears to embody a high degree of personal maturity, concern for others, and self-confidence, and to be free of any sort of narcissistic achievement drive or compulsive striving. In a functional sense, the equation may be said to identify persons likely to do well in training, as shown by evaluations near the end of such training, and at the same time to emphasize a constructive, desirable, and beneficent constellation of personal attributes.

REFERENCES

- BARRON, F. *Creativity and psychological health*. Princeton, N. J.: D. Van Nostrand, 1963.
- ERON, L. D. Use of the Rorschach method in medical student selection. *J. med. Educ.*, 1954, **29**, 35-39.
- FUNKENSTEIN, D. H. Possible contributions of psychological testing of the nonintellectual characteristics of applicants to medical school. In Helen H. Gee & J. T. Cowles (Eds.), *The appraisal of applicants to medical schools*. Evanston, Ill.: Association of American Medical Colleges, 1957. Pp. 3-27.

- GARFIELD, S. L., & WOLPIN, M. MCAT scores and continuation in medical school. *J. med. Educ.*, 1961, 36, 888-891.
- GLASER, R. Predicting achievement in medical school. *J. appl. Psychol.*, 1951, 35, 272-274.
- GOTTHEIL, E., & MILLER, CARMEN M. Prediction variables employed in research on the selection of medical students. *J. med. Educ.*, 1957, 32, 131-147.
- GOUGH, H. G. *Manual for the California Psychological Inventory*. Palo Alto, Calif.: Consulting Psychologists Press, 1957.
- GOUGH, H. G. The Adjective Check List as a personality assessment research technique. *Psychol. Rep.*, 1960, 6, 107-122.
- GOUGH, H. G., HALL, W. B., & HARRIS, R. E. Admissions procedures as forecasters of performance in medical training. *J. med. Educ.*, 1963, 38, 983-998.
- HAMMOND, K. R., & KERN, F., JR. *Teaching comprehensive medical care*. Cambridge, Mass.: Harvard Univer. Press, 1959. (Commonwealth Fund)
- HOLLAND, J. L. The prediction of college grades from the California Psychological Inventory and the Scholastic Aptitude Test. *J. educ. Psychol.*, 1959, 50, 135-142.
- HOLT, R. R., & LUBORSKY, L. *Personality patterns of psychiatrists*. New York: Basic Books, 1958.
- KEIMOWITZ, R. I., & ANSBACHER, H. L. Personality and achievement in mathematics. *J. indiv. Psychol.*, 1960, 16, 84-87.
- KELLY, E. L. A critique of the interview. In Helen H. Gee & J. T. Cowles (Eds.), *The appraisal of applicants to medical schools*. Evanston, Ill.: Association of American Medical Colleges, 1957. Pp. 78-84. (a)
- KELLY, E. L. Multiple criteria of education and their implications for selection. In Helen H. Gee & J. T. Cowles (Eds.), *The appraisal of applicants to medical schools*. Evanston, Ill.: Association of American Medical Colleges, 1957. Pp. 185-196. (b)
- KNEHR, C. A., & KOHL, R. N. MMPI screening of entering medical students. *J. Psychol.*, 1959, 47, 297-304.
- MACKINNON, D. W. The personality correlates of creativity: A study of American architects. In G. H. Nielsen (Ed.), *Proceedings of the XIV International Congress of Applied Psychology, Copenhagen, 1961*. Vol. 2. Copenhagen: Munksgaard, 1962. Pp. 11-39.
- MAXWELL, MARTHA J. An analysis of the California Psychological Inventory and the American Council on Education psychological tests as predictors of success in different college curricula. *Amer. Psychologist*, 1960, 15, 425. (Abstract)
- RICHARDS, J. M., JR., & TAYLOR, C. W. Predicting academic achievement in a college of medicine from grades, test scores, interviews, and ratings. *Educ. psychol. Measmt.*, 1961, 21, 987-994.
- RICHARDS, J. M., JR., TAYLOR, C. W., & PRICE, P. B. The prediction of medical intern performance. *J. appl. Psychol.*, 1962, 46, 142-146.
- SCHOFIELD, W. A study of medical students with the MMPI: III. Personality and academic success. *J. appl. Psychol.*, 1953, 37, 47-52.
- STUIT, D. B. Prediction of scholastic success in a college of medicine. *Educ. psychol. Measmt.*, 1941, 1, 77-84.
- SCHWARTZMAN, A. E., HUNTER, R. C. A., & LOHRENZ, J. G. Factors related to medical school achievement. *J. med. Educ.*, 1962, 37, 749-759.
- WANTMAN, M. J. Review of Medical College Admission Test. In O. K. Buros (Ed.), *Fourth mental measurements yearbook*. Highland Park, N. J.: Gryphon Press, 1953. Pp. 816-819.
- WESMAN, A. G. Review of Medical College Admission Test. In O. K. Buros (Ed.), *Fifth mental measurements yearbook*. Highland Park, N. J.: Gryphon Press, 1959. Pp. 936-938.

(Received June 24, 1963)

THE EFFECT OF PARTICIPATORY AND SUPERVISORY LEADERSHIP ON GROUP CREATIVITY¹

LYNN R. ANDERSON AND FRED E. FIEDLER

University of Illinois

This study compares the creativity of 4-man groups under 2 conditions of leadership. The Ss were 90 freshman and sophomore Navy ROTC midshipmen and 30 NROTC seniors who served as group leaders. In 15 of the groups, the leaders acted as chairmen who directed the group discussions and contributed to the task solution. In the other 15 groups, leaders acted in a supervisory capacity: they directed and guided group discussion and they were allowed to encourage members or to reject ideas, but the leaders were prohibited from contributing to the solution of the task. Groups having participatory leaders were generally superior in quantity of output while groups under supervisory leaders were superior in the quality of the product. Although leaders in the 2 conditions did not differ in their satisfaction with the group product, the participatory leaders were more satisfied with their own individual contribution to the task. The leadership styles did not produce differences in the members' esteem for the leader or in the members' morale and satisfaction with the task. Differences were found in the influence of the leader intelligence and ability scores on group creativity.

The present study compares two types of leadership conditions; namely, the more usual "participatory" situation in which the chairman of the group contributes freely to the group discussion, and a "supervisory" condition in which the leader acts as the coordinator and evaluator of the group's problem solving attempts.

The supervisory condition of leadership, although here artificially created, is not at all uncommon in real life. A director of a research program, or the supervisor of a project, frequently assigns problem solving and development tasks to a group of subordinates who then report back to him at appropriate intervals to obtain further guidance and critique. This latter experimental condition is, thus, also related to some extent to the studies by Parnes and Meadow (1959) and Meadow,

Parnes, and Reese (1959) which have shown that the creative process in groups is enhanced if the phases of idea generation and idea evaluation are separated in time. We have sought here to separate these two aspects of the creative process in the supervisory leadership condition by assigning them to different members of the group, i.e., the leader's responsibility was mainly one of evaluation, while only the members were responsible for the generation of creative ideas.

There has been very little research on participatory and supervisory leadership. Most notable is a study of small group discussion by Hare (1953) which showed that participatory leaders had more influence than supervisory leaders, and that they were also more effective in increasing intragroup agreement.

As suggested by Parnes' work (1959) as well as by Hare's study (1953), the quality of group performance was expected to be higher under supervisory leadership. On the other hand, groups in the participatory leadership condition were expected to generate a higher quantitative performance score due to having an additional group member contributing ideas.

The present study sought to extend these findings by comparing two types of leader-member interactions and by investigating per-

¹ This work was supported by Contract NR 177-472, Nonr 1834(36), "Group and organizational factors influencing group creativity," between the Office of Naval Research (Group Psychology Branch) and the University of Illinois (Fred E. Fiedler, Lawrence M. Stolurow, and Harry C. Triandis, Co-Investigators). It represents Technical Report No. 7. We are grateful for the assistance of Doyle W. Bishop, William Higgs, Lorand Szalay, Lewis Rambo, Nancy K. Barron, Patricia Chesebro, L. C. Anderson, Joseph E. McGrath, James W. Julian, and Willem A. T. Meuwese, as well as the staff of the Naval Science Department, University of Illinois.

formance on a wider range of creative tasks. The study was also designed to explore further the conditions in group processes under which leader or member abilities contribute optimally to group performance (Fiedler & Meuwese, 1963).

METHOD

Subjects

A total of 120 men, enrolled in the Naval Reserve Officers Training Corps (NROTC) program at the University of Illinois, participated in this study. Men in this program are selected on the basis of intelligence and aptitude tests, and they tend to be above the average college student in intelligence and academic performance.

The study was based on 30 four-man groups. The leaders were senior midshipmen enrolled in a Naval Leadership Course. The group members were freshmen and sophomores in the NROTC program. Leaders as well as group members were given to understand that the study constituted part of their regular training and that their performance would be evaluated and made part of their record.

Pretests

Intelligence tests. A short 15-item vocabulary test and a 15-item information test were administered and scored before the study. These tests, published by the Psychological Corporation,² are timed subtests of a larger general intelligence scale (*The Multi-Aptitude Test*) and are considered adequate indices of intellectual functioning for the purposes of this study.

Navy ROTC ratings. Regular assessments of senior midshipmen by their Navy instructors as well as by their peers were available. These consisted of the usual effectiveness ratings which asked each of the 30 seniors to judge each other senior on leadership ability and potential ability to perform a variety of leadership tasks. These ratings also contained friendship choices, e.g., "If you wanted to talk to someone about an important personal problem, which of your classmates would you prefer to consult?"

Pretest of creativity. All leaders were given individual tests parallel to those which the groups were to receive later on. These tests consisted of (a) writing three stories from a Thematic Apperception Test (TAT) card projected on a screen, (b) listing unusual uses for some common household objects (see Guilford, 1954), (c) listing creative solutions to a problem dealing with a summer recreation project (see Triandis, Bass, Ewen, & Mikesell, 1962), and (d) listing arguments (both pro and con) concerning a controversial military leadership award to be given to decisive war-time leadership successes

achieved by commanders acting against the order of their superiors.³

Experimental Manipulation

Immediately prior to the experiment the leaders of the groups in the participatory and supervisory condition were called aside and given separate instructions.

The 15 leaders in the *participatory condition* were told that they were to be chairmen of their groups, and that it was their responsibility to obtain maximum task performance. They might do this in any way they wished, including, of course, active participation in the group discussion.

The 15 leaders in the *supervisory condition* were instructed to "take a role" which they would frequently have to assume later as Navy officers, i.e., they would supervise the work of others, but they would not actually do the work itself: they could encourage their group members, they could make procedural suggestions, and they could praise or reject ideas which were proposed by their group members, but *under no circumstances* were they to contribute ideas of their own to the solutions of the problems.

Since the group sessions were part of the NROTC course, Navy officers were present in uniform throughout the testing sessions. The presence of these officers plus the fact that the leaders knew they would be evaluated on their performance served to increase the leaders' motivation to have their groups perform well.

Postsession Measures

Immediately following the group sessions, all subjects (Ss) were given individual questionnaires. These were designed to assess the success of the experimental manipulation and to indicate the reaction of the members to the group tasks and to each other. The Ss were assured that these data were to be treated confidentially.

Group session evaluations. These were obtained by means of an eight-question, five-point scale asking S how he liked his own group and how well he enjoyed the task. For example, the first two items read, "How much did you enjoy being a member of this group?" and "How much did you influence your group's product?"

Group atmosphere scale. All Ss completed a 17-item bipolar rating scale describing the climate of their group. The items contained such polar adjectives as "friendly-unfriendly"; "pleasant-unpleasant" and was scored on an 8-step scale of the semantic differential form (see Fiedler, 1962).

Effectiveness of Experimental Manipulation

Two items were included in the postmeeting questionnaire to determine the extent to which leaders had followed the experimental instructions

² Copyright by the Psychological Corporation, New York 17, New York, 1955.

³ This task and the parallel group task were developed by Dr. Lorand Szalay for this study.

regarding their role as participatory or supervisory leaders. These two items read: "Did your leader participate in the solution of the problems?", and "Did your leader add ideas to the stories, arguments, and problem solutions?" A comparison of responses of group members showed highly significant differences between the two conditions, and practically no overlap between the two groups ($p < .001$). These results, as well as postsession questionnaire responses obtained from leaders, indicate quite clearly that the experimental manipulation was successful.

CREATIVITY TASKS

Four different types of tasks were utilized to determine the generality of previously obtained findings. Each of these tasks was timed, and all groups worked on them in the same order.

TAT stories. This task consisted of writing two short stories based on TAT Card No. 11. Groups were given 15 minutes in which to complete this assignment. The stories were to be written out, and handed in at the end of the allotted time. This task had been used in previous studies of group creativity and it was here included, in part, as a marker variable, which would provide comparisons with these previous results.

Evaluation of the stories was based on a manual by which four independent judges rated five different categories on a six-point scale (Fiedler, Meuwese & Oonk, 1961). Categories to be rated were (a) originality of title, (b) originality of plot, (c) plot elaboration, (d) expressiveness of language, and (e) suspense and humor. The average intercorrelation among judges was .72. The total qualitative score represents the sum of the ratings on the five categories given to the two stories by the four judges. The correlation of ratings on the two stories for the 30 groups was .43 ($p < .05$). A simple quantitative score was based on the total number of words produced by the group on the two stories. This quantitative score was not, however, independent of the qualitative score as shown by the high correlation ($r = .65$, $p < .01$) between the two scores.

Unusual Uses test. A modification of Guilford's Test (1954) was given with instructions that the groups were to think of unusual uses for two common objects, viz., a wire clothes hanger and a ruler. Ten minutes were allowed for this task.

This task was scored on the basis of how frequently a given "use" or response occurred in any of the 30 groups. Each response was scored from one point (frequent response) to five points (unusual, off-beat or infrequent response) based on a frequency distribution of the occurrence of all of the responses produced by the 30 groups. A repetition of the same response in the same group was scored zero. The total group score represents the sum of the points given to each of the 10 uses which the group listed for each of the two items. The correlation of the two items over the 30 groups was .60 ($p < .01$). Since all but two of the groups completed the answer sheet by providing 10 uses for each item, a quantitative score could not be computed for this task. The obtained score is probably best seen as a qualitative score in that it indicates the degree to which the group was able to produce unusual responses in comparison with the other groups in the experiment.

Argument construction. A third task required the Ss to develop arguments for, as well as against, a controversial military issue. The problem involved the use of a military training program of a very rigorous and dangerous nature which could be expected to result in a relatively high rate of casualties during the training phase, but which would pay off by leading to a relatively low death rate during actual battle conditions.

After a 5-minute period in which the group members worked individually on the problem, group members were given another 5 minutes to develop additional arguments jointly, i.e., arguments which differed from those produced by the members individually. The task score was computed by rating the quality of each argument produced by the group and by the individual members on a scale from zero to three points and then summing the ratings. A quantitative score was based on the total number of arguments produced by the group and by the individual members.

Fame and immortality task. The last task was somewhat similar to the third in that the members first worked alone for 5 minutes and then as a group for the remaining 5 minutes. The problem of the task was, "how can a

TABLE 1
INTERCORRELATION OF CRITERION SCORES

	1	2	3	4	5	6	7
Supervisory condition ^a							
1. TAT quantitative		65*	−16	33	24	56*	−11
2. TAT qualitative			−40	46	08	58*	−35
3. Unusual uses qualitative				−06	26	−41	35
4. Argument construction quantitative					−26	60*	−17
5. Argument construction qualitative						−07	23
6. Fame quantitative							21
7. Fame qualitative							
Participatory condition ^a							
1. TAT quantitative		64*	09	28	10	11	−25
2. TAT qualitative			02	12	−04	23	15
3. Unusual uses qualitative				19	−46	−66*	−06
4. Argument construction quantitative					−49	−12	15
5. Argument construction qualitative						24	−03
6. Fame quantitative							05
7. Fame qualitative							

^a N = 15.
* $p \leq .05$.

person of average ability achieve fame and immortality even though he does not possess any particular talents?" This task was developed by Triandis, et al., (1962) and was scored by means of Triandis' manual. The qualitative score was the average evaluation of each of the suggested answers while the quantitative score was a tally of solutions produced. The split-half correlation of the qualitative score for the 30 groups was .89 ($p < .001$).

As shown by Table 1, the scores from different tasks are essentially uncorrelated. The relations between the Fame quality scores and TAT quantity and Fame quantity scores in the supervisory condition, and the negative relations between the Fame quantity and the Unusual Uses (quality) scores seem to reflect condition as well as task differences.

RESULTS

Supervisory versus Participatory Leadership

One of the most important questions asked by this study concerns the comparative effectiveness of the two types of leadership. The results related to this question are presented in Table 2.

As can be seen, the differences on the TAT task were nonsignificant for both the qualita-

tive and the quantitative scores. On the other hand, the groups working under supervisory leadership yielded significantly better qualitative scores on the Unusual Uses and the Fame and Immortality tasks. The participatory leadership condition led to better quantitative scores on the Argument Construction task.

These results are generally in line with the theoretical expectations based on the work of

TABLE 2
MEAN GROUP PERFORMANCE SCORES FOR GROUPS IN PARTICIPATORY AND SUPERVISORY LEADERSHIP CONDITIONS

Task	Leadership condition		<i>p</i>
	Participatory	Supervisory	
TAT stories			
Quantitative	11.60	11.00	<i>ns</i>
Qualitative	100.67	95.80	<i>ns</i>
Unusual uses			
Quantitative	—	—	
Qualitative	44.00	49.60	< .05
Argument construction			
Quantitative	29.87	27.20	< .01
Qualitative	35.20	34.00	<i>ns</i>
Fame problem			
Quantitative	4.73	4.00	<i>ns</i>
Qualitative	2.68	3.91	< .05

Parnes (1959) and others. The participatory leader was an extra source of ideas in his group and he was therefore likely to increase the quantitative output of his team. The supervisory leader, on the other hand, could devote his attention to guiding the group and to screening out ideas of low quality.

Leaders' Reactions to the Experimental Conditions

The mean scores of the leaders on the postmeeting questionnaire are presented in Table 3. The three items which discriminated between the two leadership conditions indicated that the participating leader felt he had more influence in the group (Item No. 2), was more relaxed (Item No. 3), and was more satisfied with his own *individual* performance (Item No. 4). However, there were no significant differences between the participating and supervising leaders regarding their satisfaction with the *group product* (Item No. 6). These differences on Items 2, 3, and 4 are probably

not a result of motivational differences between the two conditions since leaders in both conditions were equally interested in the tasks (Item No. 5) and they also equally enjoyed being a member of their group (Item No. 1). The fact that the supervising leaders tended to perceive their role as less influential than did leaders who participated directly in the solution of the problems may also explain why these supervisory leaders felt less relaxed in their role than the participating leaders.

A similar analysis of the eight postmeeting questionnaire items for the group members showed no significant differences between conditions on any of the eight items. This negative result is interesting inasmuch as it indicates that the experimental manipulations and the structuring of the groups had very little or no effect on members' satisfaction, tenseness, influence, or enjoyment. Similarly, no differences were found between conditions on the members' rating of the group atmosphere or the members' evaluation of their

TABLE 3
MEAN SCORE ON THE VARIOUS POSTMEETING INSTRUMENTS

Task	Leadership condition		<i>p</i>
	Participatory	Supervisory	
Postmeeting questionnaire (leaders only)			
1. How much did you enjoy being a member of this group?	3.60	3.80	<i>ns</i>
2. How much did you influence your group's product?	3.67	3.00	< .05
3. How relaxed and comfortable did you feel in this group?	4.33	3.53	< .01
4. How satisfied are you with your individual contribution?	3.40	2.73	< .05
5. Did you find these tasks interesting?	3.20	3.06	<i>ns</i>
6. How satisfied are you, personally, with your group's products?	3.93	3.60	<i>ns</i>
7. Did you feel anxious or tense in this group?	1.53	1.53	<i>ns</i>
8. How well did group members seem to communicate with each other, to understand each other's points?	3.93	3.80	<i>ns</i>
Evaluation of the "group atmosphere"			
By the leader	110.60	107.53	<i>ns</i>
By the members	108.31	107.62	<i>ns</i>
Esteem given to the leader by the group members	110.44	108.67	<i>ns</i>

Note.—A high score indicates high agreement with the item, high evaluation of group atmosphere and high esteem of leader.

TABLE 4
CORRELATION OF GROUP CREATIVITY WITH THE LEADER'S VERBAL INTELLIGENCE SCORE AND
WITH THE MEAN OF THE GROUP MEMBERS' VERBAL INTELLIGENCE SCORES

Task	Leadership condition			
	Participatory ^a		Supervisory ^a	
	Leader's IQ	Mean of members' IQ	Leader's IQ	Mean of members' IQ
TAT stories	.61**	-.48	-.01	-.59*
Unusual uses	-.40	.14	-.16	.65**
Argument construction	-.19	-.04	.65**	-.18
Fame problem	.25	.22	.03	.36

^a N = 15.
* $p \leq .05$.
** $p \leq .01$.

leader. Although major differences were found between the creativity of groups under the two leadership styles, the morale and satisfaction of the members were apparently not affected by the experimental manipulations.

The Contribution of the Leader to Group Performance

Verbal intelligence. A recent report by Fiedler and Meuwese (1963) showed that the leader's intelligence contributed to group task performance only in groups in which he was accepted by his members or where the group was cohesive. The present study provided an opportunity to explore how the contribution of the leader was affected by our experimentally introduced leader-follower interactions.

Table 4 presents the correlation between the leader's verbal intelligence and the qualitative score on each of the four tasks.⁴ It is apparent that the leader's contribution, or his ability to contribute to the task, depends in a large measure on the task itself. Thus, leader intelligence correlated with TAT performance in the participatory condition, but it was uncorrelated, if not indeed negatively

related, to group productivity on the Unusual Uses test.

We must here take note of the fact that the TAT story development required a high degree of verbal facility. The participatory leader apparently was able to integrate the flow of *his own* ideas and the ideas contributed by the group members into a consistent and unified plot. The Unusual Uses test, on the other hand, required discreet or isolated responses of high quality. On this task the verbally facile leader was likely to make many suggestions which, in the absence of adequate screening, tended to be of lower quality.

In the supervisory condition the leader's verbal intelligence is significantly related to performance only on the Argument Construction task. This task required highly logical and practical solutions rather than extremely unusual or imaginative responses from the group members. Assuming that the intelligent leader is also somewhat more "practical minded," this task then provided an opportunity for him to assess and screen the members' contributions on a more familiar criterion.

Correlations of the sum of the members' verbal intelligence scores and group performance are also presented in Table 4. It should be noted that the significant correlations with the TAT task and the Unusual Uses task were both found in the supervisory condition. This

⁴ The performance scores presented in the remaining sections of the paper are the *qualitative* scores derived for each of the four creativity tasks. This score was adopted as the single criterion score because of its closer approximation to current definitions of "creativity," while the quantitative score seems to correspond to what has been called "originality" (Maltzman, 1960).

is to be expected since the members' individual abilities should be more important when the leader cannot contribute directly to the solution of the problem. This effect is especially emphasized on these first two tasks when we consider the zero-order correlations of the supervising leader's verbal intelligence and group performance on the same tasks. The negative correlations obtained on the TAT task are of particular interest. They suggest that group members who were especially high in verbal ability produced too many unusual ideas which could not effectively be integrated into one story plot by the leader. Hence, confusion resulted with a correspondingly low group product. However, when these many unusual ideas of the members could be individually utilized without having to be integrated into an overall plot, the highly verbal members produced a better group product. This is seen most clearly on the Unusual Uses task.

A further comparison was made of the group performance under the two conditions when the verbal intelligence of the leader exceeded (or was less than) the intelligence of his group members. For each leader a difference score was computed between his own verbal score and that of his group member with the highest verbal intelligence score. Thus, a high positive "D-score" indicates that the leader was quite high in verbal intelligence when compared to his most intelligent group member, while a negative D-score indicates

TABLE 5

PARTIAL CORRELATIONS OF "D-SCORE" OF LEADERS' INTELLIGENCE (LEADER IQ MINUS HIGHEST MEMBERS' IQ) AND GROUP PERFORMANCE, WITH LEVEL OF GROUP IQ HELD CONSTANT

Task	Leadership condition	
	Participatory	Supervisory
TAT stories	.73**	.16
Unusual uses	-.28	-.89**
Argument construction	.56*	.11
Fame problem	.32	-.32

* $p \leq .05$.

** $p \leq .01$.

TABLE 6

CORRELATION OF THE LEADER'S INDIVIDUAL PRETEST SCORE ON EACH TASK AND HIS GROUP'S PERFORMANCE ON THAT TASK

Task	Leadership condition	
	Participatory	Supervisory
TAT stories	.35	-.38
Unusual uses	.43	.05
Argument construction	-.39	.11
Fame problem	.10	.08

Note.—No correlation was significant.

that the leader was actually less intelligent than his brightest group member. Since a correlation of his D-score and group performance would not account for differences in the absolute level of intelligence of the group or leader, a partial correlation was computed with the level of group intelligence (based on the highest member's score) held constant. These partial correlations are presented in Table 5. The significant positive correlations in the participatory conditions indicate that higher group creativity resulted on two of the four tasks when the leader was relatively brighter than his group members. There was no relationship of the leader's D-score and group performance in the supervisory condition except on the Unusual Uses task. On this particular task the leader who was more intelligent than his group members was actually detrimental to group creativity ($p < .01$).

The overall results from this analysis suggest some interesting differences in abilities which are desirable for effective performance as a participating and supervising leader. It would appear that the participating leader was generally more effective when he was more intelligent than his group members. The supervising leader, however, was apparently less effective on some tasks when he was more intelligent than his group members. This was particularly true on the Unusual Uses task. Here the supervisory leader's relatively greater intelligence might have hindered group creativity because the leader was too practical or critical in accepting wildly unusual and

TABLE 7
CORRELATION OF THE LEADER'S APTITUDE SCORES,
ABILITY SCORE, PEER RATINGS, AND GROUP PER-
FORMANCE

Aptitude	Leadership condition	
	Participatory	Supervisory
Leader's Military Aptitude Ratings (by NROTC instructors)		
Tat stories	.19	.13
Unusual uses	.24	-.12
Argument construction	-.08	.09
Fame problem	-.37	.26
Leader's NROTC "Buddy Ratings" (by NROTC classmates)		
TAT stories	-.17	.01
Unusual uses	.40	-.43
Argument construction	-.37	.03
Fame problem	-.48	.34
Mechanical Aptitude (NROTC battery)		
TAT stories	-.25	.26
Unusual uses	.21	-.60*
Argument construction	-.34	-.26
Fame problem	-.17	-.45
Mathematical Aptitude (NROTC battery)		
TAT stories	-.09	.32
Unusual uses	.15	-.38
Argument construction	-.13	-.04
Fame problem	-.02	-.32
Verbal Aptitude (NROTC battery)		
TAT stories	.15	.13
Unusual uses	.01	-.38
Argument construction	-.61*	.57*
Fame problem	.16	-.10

* $p \leq .05$.

unique uses which received a higher number of points.

Individual creativity tests and group performance. Since each of the 30 leaders had been pretested on tasks which were basically identical or parallel to the four group creativity tasks it seemed reasonable to expect that the scores of the leaders should correlate even more highly with group performance than the more general verbal intelligence measures. This, however, was not the case (Table 6). While the relations under the participatory leadership condition tended to be slightly higher, none of the correlations was significant, and the difference between the two conditions was very slight as far as these relations are concerned. Whether or not these lower relations merely reflect lower reliability of the individual creativity tests is a pos-

sibility which needs to be considered in interpreting these results.

Navy ROTC ratings. Table 7 shows the correlations of group performance and the sum of the "buddy ratings" choices received by the leader from each of the other senior midshipmen. Table 7 also shows the correlation of group performance and the evaluations of the 30 leaders made by their NROTC instructors (Military Aptitude Score). Although there was a high correlation of the buddy ratings and the instructor's rating ($r = .65$, $p < .05$) it appears that neither of these scores was an adequate predictor of leadership effectiveness on the creative tasks.

Additional data were also available for the 30 leaders from their aptitude testing in the Naval ROTC program. Included in Table 7 are the correlations of the various aptitude scores and group performance. Here again there appears to be very little relation of the individual aptitude of the leader and group performance. Approximately half of the correlations presented in this table are negative, while about half are positive, and only two correlations were significant. This overall lack of relationship between the leader's military ability and aptitude and the creative performance of his group points once again to the dangers in generalizing leadership results across situations and task conditions.

DISCUSSION

The results of the present study indicate the very complex relationship between type of leadership and group effectiveness. An understanding of these complexities is central not only to further theoretical development in the study of social relations, but it is also crucial to effective construction of work groups and the identification and training of successful group leaders. More and more the individual finds his occupational as well as his leisure time activities to be part of a team or group effort. His own contributions are measured by his advancement of group goals and, consequently, his personal satisfactions and rewards are distributed according to group accomplishments. The role of the leader thus becomes important in relation to both the final group achievement and the satisfaction and morale of the individual members.

Indeed, these two separate leadership functions may represent our best criteria to assess the relative effectiveness of the leader's performance.

The two "styles" of leadership which were selected for study probably represent extremes on one important dimension of leader behavior, viz., the leader's contribution to the group task. General conceptions of leadership are most likely to denote the leader as the most influential member working on the group task. This corresponds to our participatory condition where the leader was allowed to offer his own ideas in the solution of the creative tasks. The polar opposite of this type of leadership is represented by the leader whose responsibilities are solely those of coordination and supervision of subordinate's activities. Such situations are encountered in research, military and industrial settings in which a supervisory or administrative director assigns topics or problems to various task units and coordinates activity among these units but does not himself directly engage in the problem solving efforts of any one unit. Our supervisory leadership condition would seem to parallel this type of organizational structure.

Although most of the results were specific to the creative task, some tentative conclusions can be drawn regarding differences in group performance under participating and supervising leaders. Participatory groups seemed to produce a high quantity of solutions (Argument Construction task) probably because they had one more man who contributed ideas. The supervisory groups, on the other hand, seemed to produce a high quality of group solutions (Unusual Uses task and Fame and Immortality task) possibly as a result of the screening or censoring function of the supervising leader role.

The relation of the leader's individual attributes (i.e., intelligence, attitudes, and special creative aptitudes) to effective group performance was again specific to the group task as well as to the experimental condition. In general, it would appear that these personality characteristics of the leader were more highly relevant to group achievement in the participatory condition. The leader-member interaction here was apparently less formal and perhaps based more on personal

relationships. The personality traits of the supervisory leader, on the other hand, seemed to be somewhat less decisive to group creativity than the prescribed demands of his more formal and structured supervisory role.

The study also points to the importance of more carefully identifying the nature of the group task itself. Too little attention and research effort has been given to the structure and organization which the task imposes upon the group. Conversely, it becomes imperative to know those leadership styles and group structures which are most conducive to the solution of a specific task. Such an analysis has been proposed by Fiedler (1963) in a recent paper. Further research exploring the dimension of task structure is now in progress. The four group creativity tasks included in the present experiment exemplify the problems encountered in trying to generalize specific effects of leadership and group structural variables to a wide range of tasks. The results of the present study only confirm the belief that the task domain obviously remains one of the most basic of areas to be methodologically explored and systematized.

An adequate formulation of leadership behavior would, or necessity then, include a statement identifying leader abilities requisite for effective performance in a close and personal participatory role as distinguished from those abilities which bolster an individual's effectiveness in the more formal structure of the supervisory role. Our data show the danger of oversimplifying the leadership problem when these differences in group structure are ignored.

REFERENCES

- FIEDLER, F. E. Leader attitudes, group climate, and group creativity. *J. abnorm. soc. Psychol.*, 1962, **65**, 308-318.
- FIEDLER, F. E. A contingency model for the prediction of leadership effectiveness. Technical Report No. 10, 1963, University of Illinois, Group Effectiveness Research Laboratory.
- FIEDLER, F. E., BASS, A. R., & FIEDLER, JUDITH M. The leader's perception of co-workers, group climate, and group creativity: A cross-validation. Technical Report No. 1, 1961, University of Illinois, Group Effectiveness Research Laboratory.
- FIEDLER, F. E., LONDON, P., & NEMO, R. S. Hypnotically induced leader attitudes and group creativity. Technical Report No. 11, 1961, University of Illinois, Group Effectiveness Research Laboratory.

- FIEDLER, F. E., & MEUWESE, W. A. T. Leader's contribution to task performance in cohesive and uncohesive groups. *J. abnorm. soc. Psychol.*, 1963, **67**, 83-87.
- FIEDLER, F. E., MEUWESE, W. A. T., & OONK, SOPHIE. An exploratory study of group creativity in laboratory tasks. *Acta psychol.*, Amsterdam, 1961, **18**, 100-119.
- GOLB, ELLEN F., & FIEDLER, F. E. A note on psychological attributes related to the score assumed similarity between opposites (ASo). Technical Report No. 12, 1955, University of Illinois, Group Effectiveness Research Laboratory.
- GUILFORD, J. P. Factors in problem-solving. *ARTC Instructors J.*, 1954, **4**, 197-204.
- HARE, A. P. Small group discussions with participatory and supervisory leadership. *J. abnorm. soc. Psychol.*, 1953, **48**, 273-275.
- MALTZMAN, I. On the training of originality. *Psychol. Rev.*, 1960, **67**, 229-242.
- MEADOW, A., PARNES, S. J., & REESE, H. Influence of instructions and problem sequence on a creative problem solving test. *J. appl. Psychol.*, 1959, **43**, 413-416.
- PARNES, S. J., & MEADOW, A. Effects of "brainstorming" instructions on creative problem solving by trained and untrained subjects. *J. educ. Psychol.*, 1959, **50**, 171-176.
- TRIANDIS, H. C., BASS, A. R., EWEN, R. B., & MIKESELL, ELEANOR H. Team creativity as a function of the creativity of the members. Technical Report No. 6, 1962, University of Illinois, Group Effectiveness Research Laboratory.

(Received July 2, 1963)

SCALING PREFERENCES FOR TELEVISION SHOWS

FRANK J. DUDEK AND EVELYN THOMAN

University of Nebraska

The constant-sum method of psychological scaling was applied to the problem of determining scale values for the dimension of liking for 18 television shows. 2 groups ($N=376$ and 384) of relatively homogeneous Ss were employed. Groups came from different geographic areas. Stability of stimulus scale values is demonstrated when each group is divided into 2 samples and scales determined for each sample. Scales from the 2 groups are compared, and correlations are determined between CSM scale values and commercial survey ratings.

This study reports results from applying the constant-sum scaling procedure (Baker & Dudek, 1955; Comrey, 1950; Metfessel, 1947) to measure degree of liking of television shows. In view of the highly individualistic nature of the dimension of 'liking' one might question whether it is possible to demonstrate any consistency at all from one sample to another for such a dimension. This was one of the major questions explored by this study.

The usual audience surveys attempt to estimate the relative size of the viewing audience at a particular time; but a singular exception to this approach is represented by the *TvQ* ratings determined by the Home Testing Institute, Inc. (1958-1963). *TvQ* ratings are designed to measure program quality in terms of an index that shows, for the group of viewers familiar with the show, the percentage saying that it is "one of my favorites." Descriptions of the technique and various ways in which it has been applied may also be found in various trade publications (e.g., *Sponsor*, 1961, 1962; *Printers' Ink*, 1958). It was of interest to compare scale values obtained in this study with other survey results.

METHOD

Scaling method. The constant-sum method has been utilized with some success in connection with the scaling of various kinds of subjective and psychological dimensions (Baker & Dudek, 1955, 1957; Dudek & Baker, 1956). Details of the procedure are reported by Baker and Dudek (1955, pp. 296-297). Essentially, the constant-sum method requires the subject (*S*) to divide 100 points between two stimuli to indicate their relative magnitudes with respect to some specified dimension. Judgments made in this

way presumably convey somewhat more information than is provided by a simple preference statement. The task is one that Ss can easily understand and which they find relatively easy to carry out even for subjective variables.

Stimuli. The large number and variety of shows available as potential stimuli poses some problems. Since the constant-sum method is essentially a "paired-comparison" procedure, obtaining complete judgments on even a modest number of stimuli would require a large number of judgments. For this study 18 shows were selected as stimuli to be scaled. They were: (1) *Danny Thomas*, (2) *Andy Griffith*, (3) *Stump the Stars*, (4) *The Price is Right*, (5) *David Brinkley's Journal*, (6) *Rifleman*, (7) *Dobie Gillis*, (8) *Beverly Hillbillies*, (9) *Third Man*, (10) *Dick Van Dyke*, (11) *Our Man Higgins*, (12) *Hazel*, (13) *Leave it to Beaver*, (14) *McHale's Navy*, (15) *Jack Benny*, (16) *Joey Bishop*, (17) *Red Skelton*, (18) *Jackie Gleason*.

With one exception all of the shows are comedies; and all but two are half-hour shows. All are televised during 'prime time' (i.e., between 6:30 and 10:00 P.M.).

If every stimulus were to be paired with every other one this would require 153 judgment items. By grouping the shows into four sets, and having certain shows overlap at least two sets, it was possible to reduce the number of items necessary to get information about all shows to 57. The four sets of shows could be characterized as: (a) six half-hour shows telecast on Monday nights; (b) six situation comedies televised during the week and having as a title the series name or a character name; (c) five half-hour shows telecast on Wednesday night; and (d) seven shows telecast throughout the week, where the title of the show was the same as the male star's name. Shows belonging to each 'set' and the overlap of sets are indicated in Table 1.

Subjects. The Ss consisted of 376 students from the University of Omaha¹ and 384 students from the University of Nebraska. All were enrolled in undergraduate psychology courses. Apart from the fact

¹ The authors would like to thank Francis M. Hurst and William Jaynes of the University of Omaha for their assistance in obtaining the Omaha data.

that these Ss were readily available, another consideration was involved in choosing college students as Ss. It was felt that if any degree of agreement in scales could be demonstrated at all, it would be most likely for groups that are relatively homogeneous. If little consistency in scale values is found even for fairly homogeneous groups one might argue that preferences for stimuli of this kind are so variable and idiosyncratic that little scaling is possible.

Experimental procedure. Judgments were obtained by having Ss mark prepared booklets presenting the pairs of stimuli to be judged. All judgments were obtained in group sessions, and these occurred during a single week.

The first page of the questionnaire consisted of the names of 30 television shows (12 titles in addition to the 18 mentioned above). For these items, S simply indicated whether or not he was familiar with the show. Succeeding pages consisted of paired items involving the 18 shows for which scale values were determined. The Ss indicated their relative degree of liking for each show in a pair by dividing 100 points between the members of each pair. Instructions reminded them that point assignments implied 'ratios' and that a judgment like 75-25, for example, implied that they like one show "three times as much" as the other. In general an attempt was made to convey the idea that the actual point assignments should reflect relative magnitudes. The Ss were instructed to omit items involving a show with which they were unfamiliar.

Scale values were determined for the stimuli within each set of stimuli by the procedures described by Baker and Dudek (1955, p. 297). To determine scale values for all items in terms of a common unit, values from stimulus sets *b* and *c* were first combined (Stimulus 11 was considered as the unit of the scale). Then on the basis of Stimulus 10, common to sets *c* and *d*, values were determined for the three sets, and finally on the basis of Stimuli 1 and 2, common to sets *a* and *d*, a set of values for all stimuli was determined where the unit stimulus was Stimulus 11.

RESULTS AND DISCUSSION

Table 1 presents the basic information of the study. Of primary interest is the stability of scale values determined for various samples. Each of the large groups was split into two samples and scale values determined for each subsample as well as for the entire group. While the degree of agreement between samples is evident from inspecting the scale values in Table 1, the relations are shown more explicitly in Figure 1, where the points are plotted using the two sample values from each of the groups as arguments. Scale values

TABLE 1
STIMULUS SETS, PROPORTION OF FAMILIARITY AND SCALE VALUES FROM
LINCOLN AND OMAHA GROUPS

Show number	Set	P _O ^a	P _L ^b	SV _{O1} ^c	SV _{O2} ^d	SV _{L1}	SV _{L2}	SV _{TO}	SV _{TL}
1	a, d	.95	.94	1.15	1.11	.97	1.05	1.13	1.01
2	a, d	.87	.85	1.31	1.37	1.32	1.45	1.35	1.38
3	a	.38	.35	.77	.75	.77	.92	.76	.84
4	a	.90	.91	.83	.80	.78	.88	.81	.83
5	a	.66	.64	1.41	1.47	1.67	1.69	1.44	1.68
6	a	.90	.82	1.08	1.03	1.13	1.17	1.06	1.15
7	b, c	.92	.89	1.14	1.14	1.06	1.11	1.14	1.08
8	b, c	.78	.72	1.30	1.46	1.07	1.23	1.38	1.14
9	c	.57	.45	1.09	1.01	1.23	1.27	1.05	1.24
10	c, d	.74	.51	1.32	1.28	1.10	1.34	1.30	1.21
11	b, c	.44	.26	1.00	1.00	1.00	1.00	1.00	1.00
12	b	.72	.66	1.05	1.10	.90	1.06	1.07	.98
13	b	.88	.78	1.01	1.05	.90	.95	1.04	.92
14	b	.71	.46	1.47	1.68	1.31	1.37	1.58	1.33
15	d	.89	.94	.98	1.08	.99	1.16	1.03	1.07
16	d	.71	.66	1.05	1.11	.97	1.14	1.08	1.05
17	d	.94	.96	1.39	1.55	1.39	1.68	1.47	1.53
18	d	.89	.88	1.38	1.40	1.11	1.29	1.39	1.20

Note.—N = 362 for Omaha group and 384 for Lincoln Group.
^a O = Omaha.
^b L = Lincoln.
^c 1 = Sample one.
^d 2 = Sample two.

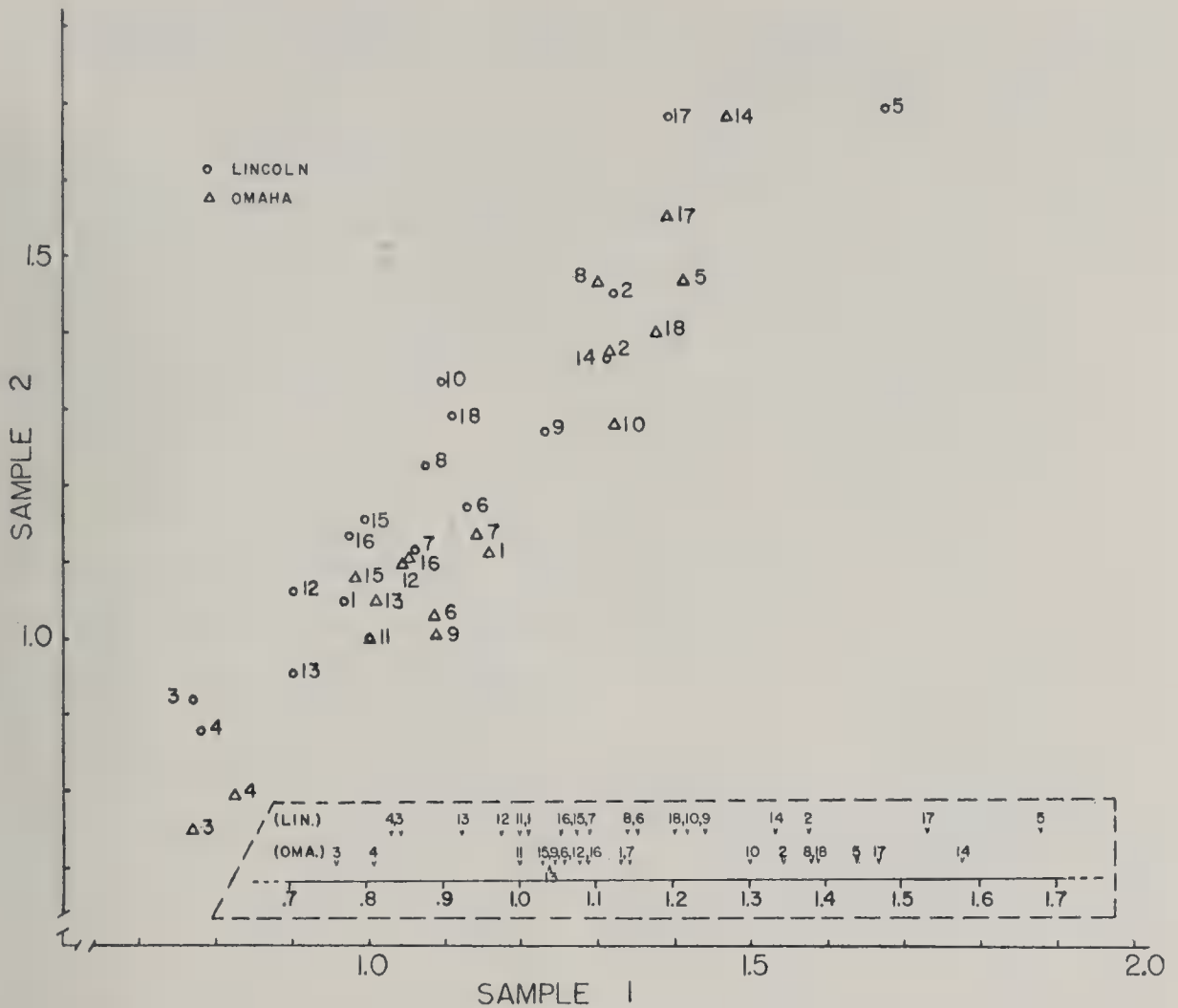


FIG. 1. Relation between stimulus scale values for two samples of each group. (Group values are shown within the insert.)

for the Omaha and Lincoln groups in terms of linear distance are shown within the insert of Figure 1. The agreement between samples is fairly high in each group, as is the agreement between the two different geographic groups. Expressed as correlation coefficients, r for the two Lincoln samples is .94, for the two Omaha samples it is .96, and the correlation between the Omaha and Lincoln groups is .82.

At the outset we questioned whether a dimension like liking of TV shows represented a scalable dimension. The results seem to indicate clearly that it is. The amount of agreement demonstrated seems even more impressive viewed against the fact that some of the comparisons were based on relatively few judgments. While the original groups were

large, it was not the case that every individual in the group could rate every item pair. As can be seen in Table 1, the proportion of individuals familiar with the various shows varied from less than 30% to more than 90%. Thus the ratios involving stimulus 11, for example, were based on approximately 40 judgments for each of the Omaha samples. For most items, however, the number of judgments available for establishing the ratios was approximately 100 for each of the Lincoln samples, and 125 for each of the Omaha samples.

While a group of college students does not represent any sort of 'typical' television audience, it may be of interest to compare the CSM scale values with audience survey ratings made in the areas from which our

samples came.² While audience ratings at different times correlate on the order of .75-.90 for the shows represented in our study, the correlation between the scale values we obtained and the audience ratings are on the order of .35-.45. Correlations between scale values and *TvQ* scores were substantially higher.³ *TvQ* scores were available for two samples (a) "adults 18 to 34 years of age," and (b) "adults 18 years and over who attended or graduated from college." Correlations between *TvQ* scores from sample (a) and scale values for the Omaha and Lincoln groups were .78 and .69, respectively. Corresponding correlations for *TvQ* scores from sample (b) were .63 and .75.

We conclude from these results that the problem of measuring a dimension like 'liking' of television shows is amenable to psychological procedures of scaling and that the scale values obtained show reasonable cor-

respondence with other indices designed to assess a somewhat comparable dimension.

REFERENCES

- BAKER, KATHERINE E., & DUDEK, F. J. Weight scales from ratio judgments and comparisons of existent weight scales. *J. exp. Psychol.*, 1955, **50**, 293-308.
- BAKER, KATHERINE E., & DUDEK, F. J. Scaling line-lengths with a modification of the constant-sum method. *Amer. J. Psychol.*, 1957, **70**, 81-86.
- COMREY, A. L. A proposed method for absolute ratio scaling. *Psychometrika*, 1950, **15**, 317-325.
- DUDEK, F. J., & BAKER, KATHERINE E. The constant-sum method applied to scaling subjective dimensions. *Amer. J. Psychol.*, 1956, **69**, 616-624.
- DUDEK, F. J. A comparison of scale values for adverbs determined by the constant-sum method and a successive intervals procedure. *Educ. psychol. Measmt.*, 1959, **19**, 539-548.
- HOME TESTING INSTITUTE, INCORPORATED. *Must reading about HTI and TvQ*. (No. 10, June 1960 through No. 24, July 1963) Manhasset, New York: HMI, 1960-63.
- METFESSEL, M. A proposal for quantitative reporting of comparative judgments. *J. Psychol.*, 1947, **24**, 229-245.
- Printers' Ink*, Q-ratings promise to help advertisers predict success or failure of TV shows. 1958, **265**, 44-45.
- Sponsor*, Can you predict TV hits and flops? 1961, **15**, 34-36.
- Sponsor*, How to predict program hits. 1962, **16**, 28-30.

(Received July 8, 1963)

² The authors would like to express their thanks to Gene Thompson of the American Research Bureau for making available to them the ARB Television Market Reports regarding the audience survey results in the Lincoln and Omaha areas.

³ The authors are indebted to Henry Brenner and Herb Altman of the Home Testing Institute, Inc. for supplying the *TvQ* data for the shows represented in their sample.

THE FACTORIAL STRUCTURE OF SELECTED CONSUMER CHOICE PARAMETERS AND THEIR RELATIONSHIP TO PERSONAL VALUES¹

JOHN R. RIZZO AND J. C. NAYLOR

Ohio State University

Importance ratings of 104 respondents of 6 choice parameters for each of 5 product types were related to value scores on the Allport-Vernon-Lindzey Scale (AVL) using (a) an analysis of variance of raw score judgments and (b) multiple regression of the AVL against factor scores obtained from a factor analysis of the ratings. The Economic scale was significantly related to choice parameters ($p < .05$) using raw judgments although none of the factored criteria dimensions (quality, cost, performance, and style) were predictable from the scale, either alone or with the inclusion of 4 personal history items as predictors. However, 3 of the zero-order correlations between value scales and style (the Social, Political, and Religious scales) were significant at $p < .05$. The first 4 factors to emerge from the respondents ratings were found to account for the majority of factor variance, although 19 factors were extracted. Most factors were clearly definable as either choice dimension or product-type factors, with the former accounting for more variance.

There have been several attempts to develop general theoretical frameworks for consumer choice behavior, the most notable of which are those of Katona (1951), Hayes (1950), Benson (1955), and Edwards (1954). While all of those mentioned take indirect cognizance of the importance of human values in choice behavior, none treat the value concept with the significance it may deserve. Similarly, examination of psychological research on consumer behavior provides much the same picture. There are, of course, numerous studies which examine such choice predictors as needs, desires, and motives. For example, Benson and Peryam (1958) found a high and significant curvilinear relationship between cost and stated preference for various meat products. Brown (1958) has shown how various types of packaging affect consumer perceptions of the freshness of bread. Mueller's results (1957) clearly demonstrate the importance of consumer attitudes on purchasing, and relationships between personality variables and product use were obtained by Tucker and Painter (1961) and by Koponen (1960). However, there is little or no empirical research directly relevant to values as predictors of consumer behavior.

In fact, although there has been a considerable amount of research concerned with human values per se (Albert & Kluckhohn, 1959; Dukes, 1955; Morris, 1956; and Thurstone, 1959) surprisingly little has been done in the area of predicting choice behavior of *any sort* from the value commitments of individuals much less consumer behavior. Several exceptions may be noted. Shepard and Bayton (1951) predicted purchases of men's suits from the values of comfort, orderliness, economy, pleasure, social approval, and recognition. Similarly Brumback (1963) found that a measure of values toward behavior in organizations could be used to predict choices of types of vacation tours as well as choices of alternatives in a two-man game situation.

It is suggested that one of the more promising approaches to an understanding of consumer behavior would be through the study of the role of values operative in the choice or consumer act. As long as one is willing to subscribe to the notion that human behavior is goal-oriented, then the concept of values must be considered a potential parameter of behavior. The present study represented an initial attempt to relate measures of values to several aspects of the consumer-choice situation. The aspects considered for study were the type of product and characteristics or dimensions associated with the pur-

¹ This investigation was supported in part by a Public Health Service fellowship (No. MPM-15,540) granted the senior author by the National Institute of Mental Health, Public Health Service.

TABLE 1—ANALYSIS OF VARIANCE COMBINED SUMMARY TABLE

Source of Variance	df	Theoretical		Economic		Esthetic		Social		Political		Religious	
		MS	F	MS	F	MS	F	MS	F	MS	F	MS	F
Between Ss	103												
AVL value													
group (C)	3	20.85	1.17	54.91	3.27*	22.32	1.26	31.40	1.80	9.38	—	11.32	—
S's/group	100	17.80		16.78		17.76		17.49		18.15		18.09	
Within Ss	3016												
Product (A)	4	168.81	87.02**	168.81	90.76**	168.81	86.57**	168.81	91.25**	168.81	87.92**	168.81	80.39**
Dimension (B)	5	331.23	66.51**	331.23	65.72**	331.23	66.65**	331.23	66.51**	331.23	66.51**	331.23	66.51**
A × B	20	55.49	38.01**	55.49	38.01**	55.49	38.01**	55.49	38.27**	55.49	38.80**	55.49	34.25**
A × C	12	2.36	1.22	4.59	2.47**	1.43	—	4.97	2.69**	2.62	1.36	2.36	1.12
B × C	15	6.98	1.40	5.31	1.05	7.42	1.49	7.30	1.47	7.38	1.48	6.95	1.40
A × B × C	60	1.51	1.03	1.47	1.00	1.28	—	1.96	1.35	2.38	1.66**	2.12	1.31
A × Ss	400	1.94		1.86		1.95		1.85		1.92		2.10	
B × Ss	500	4.98		5.04		4.97		4.98		4.98		4.98	
AB × Ss	2000	1.46		1.46		1.46		1.45		1.43		1.62	

Note.—Mean squares and *F*'s for the analysis of each Allport-Vernon Scale versus product judgment data.

* *p* < .05.

** *p* < .01.

chase of that product which would be of importance in the consumer decision.

METHOD

Subjects

One hundred four volunteer college undergraduates at the Ohio State University served as subjects (*Ss*). The sample consisted of 49 males and 55 females, with an age range from 17 to 32 years.

Procedure

Each *S* was administered (*a*) the Allport-Vernon-Lindzey Scale of Values (AVL) and (*b*) a questionnaire designed to measure the relative importance of six product dimensions in the purchase of five different products. The *Ss* were tested in groups ranging in size from 8 to 20, and each testing session lasted about 45 minutes. The *Ss* were asked to rate, using a 9-point Thurstone scale, the importance of each dimension in the purchase of each product. For example, with clothing, *Ss* were asked "when purchasing clothing, how important do you feel each of the following factors are?"

Questionnaire

The consumer-preference questionnaire consisted of four personal history items (age, sex, father's occupation, and home town size) in addition to the questions dealing with product-purchase dimensions. Five products (clothing, car, major appliance, home, and food) were used. Each product was followed by six dimension scales. These purchase dimensions were style, comfort, cost, performance, place of purchase, and manufacturer. For the product "home," the dimension of performance was qualified with the additional word "quality," the dimension of place of purchase with the qualifier "location," and the dimension of manufacturer with the qualifier "builder." For the product "food," style was qualified with the word "looks," and performance with quality. The dimension of comfort for the product of food was replaced with the dimension of "convenience."

RESULTS

Analyses of Variance

Results were analyzed employing a 4 × 5 × 6 factorial, repeated-measures analysis of variance design. One analysis of variance was computed for each of the six Allport-Vernon value scales. The three factors in each design were (*a*) Allport-Vernon value scale score (*Ss* were divided into quartiles in each analysis on the basis of values scale scores, thus creating four levels of this factor), (*b*) product type, with five levels (e.g., clothing, car, etc.), and (*c*) dimensions with six levels (e.g., style, comfort, cost, etc). Cell entries in each analy-

sis were the Ss' ratings of the importance of each dimension in the purchase of each product. Results of these analyses of variance are presented in summary in Table 1.

Regarding the main effect of value scores, Ss' mean ratings varied only as a function of their scores on the economic value scale, with the second quartile group having the highest mean, followed in order by the third, fourth, and first quartile groups. The main effect of values was insignificant for the remaining value scales.

The mean squares for the main effect of product, the main effect of dimension, and for the product by dimension interaction were identical in all six analyses of variance. This occurs because the ratings analyzed changed location in the design only as a function of Ss' quartile location on each of the value scales. The main effects of product and of dimension, and the product by dimension-interaction effect were significant at the .01 level in all analyses.

Considering first the main effect for product, mean ratings associated with this effect may be interpreted as "intensity of feeling" associated with the product. These means, in order of size, were as follows:

Product	Mean
Home.....	7.66
Car.....	7.00
Appliance.....	6.75
Food.....	6.54
Clothes.....	6.34

Mean size for product tends to vary directly with the product cost, with higher means associated with high cost. There appears to be a higher degree of personal involvement for more expensive products.

The mean ratings for the significant main effect of dimension were, again in order of size:

Dimension	Mean
Performance.....	8.01
Comfort.....	7.37
Cost.....	7.17
Style.....	6.56
Manufacturer.....	6.15
Place of purchase.....	5.87

Thus, performance was considered most important in the purchase of products, while

TABLE 2
MEANS ASSOCIATED WITH THE PRODUCT
BY DIMENSION INTERACTION

Dimension	Cloth- ing	Car	Appli- ance	Home	Food
Comfort	7.64	7.19	7.68	8.09	6.23
Style	7.42	6.84	5.33	7.30	6.04
Performance	6.96	8.36	8.40	8.22	8.11
Cost	6.65	7.47	7.28	7.61	6.80
Manufacturer	4.69	6.76	6.53	6.85	5.88
Place of purchase	4.67	5.39	5.28	7.84	6.16

manufacturer and place of purchase were considered of least relative importance.

The means associated with the product by dimension interaction indicate the relative importance of the dimensions in the purchase of each product. Table 2 lists these means by product. For clothing, comfort, and style are rated highest, followed by performance and cost, while the lowest means for clothing are for the dimensions of manufacturer and place of purchase. Manufacturer and place of purchase are also considered least important in the purchase of a car, while performance is considered most important for this product. Cost, comfort, and, perhaps somewhat surprisingly, style follow performance in importance for the car. Regarding the purchase of a major appliance, performance carries the highest mean again. As with the car, cost and comfort follow performance in importance. Manufacturer is fourth in importance for the appliance, its highest position relative to the other dimensions for any other product. Style and place of purchase are of lowest importance in appliance purchase. For the home, performance (quality) and comfort were rated highest. Place of purchase (location in this case) moves to third, the highest relative position for this dimension as compared to its position and mean size for any other product. Style and manufacturer (builder) are considered to be of least relative importance in home purchase. Performance (quality), for the fourth time out of five products, is rated highest in the purchase of food. Cost was the second highest mean for this product, followed by convenience, place of purchase, style (looks), and manufacturer in that order.

TABLE 3
CORRELATION AND RESIDUAL MATRIX FOR FACTOR

Item	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1	70	01	02	00	-01	01	-01	00	01	-03	00	01	-01	02
2	09	60	00	01	-01	-01	00	00	00	-01	00	01	00	01
3	-05	24	74	02	00	-02	-01	00	00	-01	00	02	01	00
4	-03	41	26	77	-01	01	00	00	01	-01	-01	-01	-01	00
5	15	-12	-08	-01	73	01	01	00	00	02	00	-01	00	00
6	06	-09	-07	29	46	70	01	01	-01	01	00	01	01	00
7	55	07	04	13	30	25	86	01	00	00	00	00	01	-01
8	17	48	16	29	-01	12	32	84	01	00	00	00	00	01
9	02	40	57	24	-29	-28	-03	27	86	00	00	00	00	00
10	04	40	24	33	-04	03	06	35	24	57	01	-01	01	-01
11	07	07	-03	19	32	37	23	13	-05	00	85	00	00	-01
12	15	-03	05	08	33	41	22	00	-06	07	55	83	00	00
13	18	14	19	26	13	30	40	16	10	10	25	22	74	01
14	15	30	05	30	10	17	19	44	18	26	14	14	35	72
15	08	28	62	12	-15	-16	12	23	71	24	02	-02	06	12
16	12	32	11	31	10	17	33	45	16	43	04	10	12	44
17	00	01	06	25	44	35	11	00	00	-04	76	52	30	20
18	10	08	00	33	32	46	20	08	-03	16	61	75	28	27
19	49	08	-07	17	24	34	57	25	-06	17	17	27	40	29
20	22	40	14	35	09	22	36	58	21	44	10	18	28	52
21	03	30	50	16	-06	-09	02	14	67	22	11	03	04	11
22	29	27	17	31	02	15	30	45	14	42	15	15	17	25
23	30	17	04	14	24	23	38	36	11	31	26	31	22	32
24	-05	16	-01	25	38	38	28	28	-05	15	42	50	37	21
25	15	27	10	08	12	01	21	21	14	14	15	14	38	35
26	08	19	19	17	10	19	19	30	25	01	21	13	37	17
27	13	28	47	20	-03	03	05	34	54	19	-06	00	14	22
28	25	35	02	41	14	16	25	39	16	44	26	12	18	45
29	14	01	-13	02	29	29	16	23	-06	-02	38	32	15	34
30	10	-02	-06	25	27	57	16	19	-02	10	41	50	43	31

There were two significant product by value scale interactions, these occurring for the social and economic values. An examination of the individual means associated with these interactions revealed no strongly systematic trends, rendering a detailed interpretation somewhat difficult. However, there was a slight tendency on the social value for high and low means to be associated with the two middle value quartiles, while on the economic value there was a slight tendency for low means on extreme quartiles and high means in the second quartile group.

Factor Analysis

The Ss' ratings on the six product dimensions for the five products were intercorrelated, yielding a 30 × 30 matrix. This matrix was factor analyzed employing a centroid

factor-analysis program with a varimax rotation. A total of 19 factors were extracted. Table 3 shows the intercorrelation and residual matrices. No residual exceeded a size of ±.027. The rotated factor structure is given in Table 4.

A number of the factors were relatively easy to interpret. Factor 1 is a combination of manufacturer and place of purchase loadings and appears to be a general "quality" factor. All products are represented on this factor's highest-loading items. Factor 2 is clearly a "cost" factor, with the five highest-loading items representing all of the five products in combination with the cost dimension. The third factor's highest-loading items refer primarily to "performance," with several slightly lower-loading items being comfort and place of purchase for the home and clothing com-

TABLE 3

ANALYSIS JUDGMENTS OF PRODUCT DIMENSIONS

15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
00	00	00	-01	01	01	-01	02	01	00	-01	00	-01	00	00	00
01	00	00	-01	00	00	00	00	00	01	01	00	-01	-01	00	00
01	00	00	-02	-01	00	00	00	00	01	00	00	-01	-01	01	-01
00	00	01	00	01	00	-01	00	01	00	-01	00	00	01	00	00
-01	00	01	00	01	-01	01	-01	00	01	00	00	01	01	00	00
00	00	-02	01	-02	01	02	-01	01	-01	00	00	-01	-02	00	-01
-01	-00	-01	02	-01	01	01	00	00	-01	01	-01	00	00	00	00
00	-01	00	00	00	00	00	00	00	00	-01	01	00	00	00	-01
00	00	00	00	-01	00	00	-01	00	00	00	00	00	00	00	00
-01	01	-01	03	00	01	00	00	00	-02	02	-01	01	02	-01	01
00	00	01	00	00	00	00	00	00	00	-01	00	01	00	00	00
00	-01	00	00	01	00	00	00	00	01	-01	01	00	01	00	00
00	-01	00	00	01	-01	00	00	00	00	00	01	00	01	-01	00
00	01	-01	01	-02	02	01	00	01	-01	00	00	00	00	00	-01
82	01	01	00	01	00	00	01	01	00	00	-01	01	00	00	00
20	89	01	00	01	-01	-01	00	-01	00	00	01	00	00	00	01
01	-02	89	01	01	00	-01	00	-01	00	02	-01	00	01	-01	02
-03	10	66	91	00	-01	01	00	00	00	00	00	00	-02	00	-01
11	25	11	27	82	00	00	00	-01	00	00	00	01	00	01	00
18	47	12	21	57	86	00	00	00	01	00	00	00	00	-01	00
68	14	14	13	07	18	75	00	00	-01	00	01	-01	00	-01	00
25	47	12	31	32	42	20	70	00	00	00	00	00	00	00	00
22	36	21	36	52	53	18	48	62	00	00	01	-02	-01	00	00
-09	25	38	57	28	36	-03	26	44	87	00	-01	00	00	00	01
02	18	20	19	18	31	08	09	22	27	67	00	02	00	00	00
14	30	23	19	16	25	20	24	22	26	36	64	00	00	00	00
54	12	-07	08	19	25	43	23	19	00	19	34	76	01	-01	02
20	48	21	30	37	45	25	51	48	28	32	20	23	75	01	01
-07	02	39	45	24	24	11	12	25	33	33	12	-04	31	84	01
-10	16	51	61	33	28	05	29	31	53	13	29	09	24	60	98

Note.—Above-diagonal values are residuals; below-diagonal values are Pearson r 's; and communalities are in the diagonal cells.

fort. Factor 4 is clearly a "style" factor, but only for the products of car, clothing, and home. Factor 5 is somewhat difficult to interpret. Only one very high-loading item occurs, this being for "appliance-style," an item which did not load highly on the aforementioned style factor. Manufacturer does appear somewhat consistently for the lower-loading items.

Some factors appear to be predominantly representative of products rather than of dimensions. For example, Factors 6 and 7 are primarily "food" factors. Food-cost and food-convenience reappear on Factor 11. Factor 9 is primarily "clothing" for the performance, comfort, and manufacturer dimensions, while Factor 10 is similarly "clothing" but only for the manufacturer and place of purchase dimensions. Factor 12 is clearly a "home" factor.

Regression Analysis

The first four factors extracted were the most definitive and together accounted for 52% of the total factor variance. Factor scores were computed for each S on each of these factors using the highest-loading items. For Factor 1, Items 11, 18, 17, 12, 30, 24, and 6 were used; Factor 2 used Items 15, 9, 21, 3, and 27; Factor 3 used Items 10, 22, 28, 16, 23, 20, and 2; and Factor 4 scores were comprised of scores from Items 7, 1, and 19. Thus, each person was given a quality score, a cost score, a performance score, and a style score. These scores were used as criteria and were predicted, using multiple regression, by the scores on the six value scales and the four personal history items. The personal history items of father's occupa-

TABLE 4—ROTATED FACTOR STRUCTURE FOR FACTOR ANALYSIS OF JUDGMENTS OF PRODUCT DIMENSIONS

Product	Item number	Factors																		
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
Clothing	1	03	01	13	78	-01	-09	-06	16	-07	01	-13	-03	-02	-01	02	-02	10	11	-02
	2	00	-27	-32	01	01	-24	09	-04	27	-11	00	-04	-35	00	-03	-09	36	02	03
	3	-00	-73	-12	-07	17	-09	12	03	14	06	-02	04	-06	-28	02	08	-17	-05	00
	4	17	-14	-27	-02	13	-00	00	-05	77	00	-04	-06	-10	01	-07	-08	04	02	-01
	5	30	12	01	18	00	-07	-07	-12	-05	75	01	-02	04	00	-03	-04	01	02	-02
	6	34	17	-04	07	25	15	-28	-02	24	42	-20	-17	-02	-02	-16	03	-10	-13	16
Car	7	09	-03	-02	80	16	-06	01	-18	10	13	15	-09	-16	-02	-18	-04	-15	-09	-02
	8	-03	-15	-30	15	03	-09	-12	-11	09	-03	-10	-15	-76	01	-14	-17	01	01	01
	9	-05	-82	-04	-03	01	-10	00	-07	09	-25	-07	01	-08	02	-07	-10	10	24	01
	10	-00	-19	-67	-01	02	-04	03	-04	13	-04	03	-11	-09	-11	-01	-08	08	00	12
	11	83	-01	01	08	11	-04	-07	01	04	09	12	01	-19	27	-02	03	-34	-04	08
	12	76	02	-06	15	00	-02	-12	-18	-09	06	-12	-11	10	-30	-10	-02	01	-02	16
Appliance	13	17	-08	-02	23	72	-23	-07	-13	14	01	00	-08	00	-03	-03	-14	00	-02	01
	14	10	-06	-23	07	14	-20	-18	01	12	03	-05	-15	-17	02	-12	-68	02	02	-00
	15	-02	-86	-14	11	-02	07	07	00	-05	-06	-03	-02	-06	08	-03	-07	-06	-07	04
	16	-03	-08	-49	13	-01	-05	03	-08	10	09	08	-08	-15	-02	-70	-25	02	-04	06
	17	79	-08	03	-05	18	-08	-12	06	10	26	17	-03	-00	13	03	-07	-03	16	-17
	18	80	01	-17	09	03	-05	-24	-21	16	04	-18	-04	09	-13	-01	-08	07	-07	-12
Home	19	10	00	-16	58	20	00	-13	-08	05	10	-13	-55	34	11	-02	-11	02	-12	05
	20	04	-12	-37	18	07	-18	-10	-10	15	04	02	-66	-30	-09	-11	-21	03	07	-01
	21	09	-79	-11	01	-08	-03	-14	04	03	-02	04	-05	05	12	-07	03	19	-05	-07
	22	14	-12	-63	24	08	07	-10	-03	06	-06	-13	-05	-18	04	-18	03	-09	-03	-28
	23	23	-12	-42	29	02	-06	-09	-26	-12	10	-10	-35	-08	16	-08	-08	-07	03	-06
	24	41	08	-18	00	20	-13	-16	-70	07	19	00	-13	-14	-00	-10	00	02	01	00
Food	25	10	-05	-09	10	15	-76	-09	-07	00	01	-02	-07	-06	01	-05	-13	02	00	01
	26	14	-19	03	02	32	-34	-07	-03	03	04	-24	-08	-18	11	-44	13	-01	09	-16
	27	-07	-58	-07	06	06	-12	-00	02	06	05	-58	-08	-17	-01	-01	-09	00	00	-01
	28	16	-10	-58	20	-08	-24	-16	-03	23	04	-08	-08	-05	35	-11	-22	-01	04	-01
	29	31	04	00	10	-07	-26	-75	04	-04	12	10	-06	-12	07	07	-16	-01	10	-05
	30	43	07	-12	04	38	10	-71	-13	09	08	-16	-10	-01	-06	-14	-06	00	18	04

tion and home town size were each coded into five categories for purposes of these analyses. Sex was coded 0 or 1 for female and male, respectively, and age was that number, in years reported by S.

Table 5 shows the intercorrelation matrix for the 10 predictors and four criteria used in these analyses. None of the R^2 values were found to be significant. The R^2 for the quality factor was .110; for the cost factor, .067; for the performance factor, .078; and for the style factor, .151. A regression was also computed against the style criterion using only the value scales as predictors (Table 5 shows three scales significantly correlated with style). The resulting R^2 of .127 gave an $F = 2.04$ with 6 and 89 df —not significant at $p < .05$.

DISCUSSION

It was not unexpected to find that the respondents tended to generally appear more involved in more expensive purchases than less expensive ones, as indicated by the ratings. Less expected was the finding that cost did not emerge any higher than third in terms of overall importance, being led by both the performance and comfort dimensions. Of particular interest, however, were the interaction findings which indicated the high importance of style with clothes, but the low importance of style with the products of appliance and surprisingly, car. Similarly, place of purchase was judged extremely low in the context of clothes, car, and appliance, but of medium importance with home and, again surprisingly, food, implying that the respondents might be more apt to shop for the former products than to rely on a particular known source for the product.

The use of values, at least as measured by the AVL scale, as a predictor for consumer-choice preference was certainly not fruitful. Only in the case of the relationship between the Economic scale (the most logical scale from an intuitive basis) and the raw-score judgments was a significant relationship obtained. While it was here that three scales (Social, Political, and Religious) all had significant correlations with the style factor scores, still they were not able (when combined with the other value scales) to produce predictability significantly different from

TABLE 5
INTERCORRELATION MATRIX OF PREDICTORS AND CRITERIA USED IN MULTIPLE REGRESSION ANALYSES

	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1. Quality factor	1.000	.021	.278*	.304*	-.070	-.138	-.055	.214*	.041	.007	-.027	-.016	-.128	-.056
2. Cost factor		1.000	.611*	.089	-.027	-.027	.065	.010	-.058	-.068	-.045	-.030	-.089	-.138
3. Performance factor			1.000	.539*	.023	.093	.156	-.048	.018	-.041	.146	.011	.094	.063
4. Style factor				1.000	.073	.157	.016	-.234*	.245*	-.264*	.153	.111	.052	.072
5. Theoretical value					1.000	.206*	-.141	-.378*	-.021	-.451*	.412*	.297*	-.127	.090
6. Economic value						1.000	-.406*	-.472*	.292	-.480*	.085	.273*	-.086	-.085
7. Esthetic value							1.000	-.136	-.219*	-.165	-.114	-.442*	.012	.176
8. Social value								1.000	-.411*	.337	-.164	-.081	.075	.019
9. Political value									1.000	-.399*	.042	.325*	.129	-.036
10. Religious value										1.000	-.201*	-.204*	.009	-.034
11. Age											1.000	.355*	-.041	-.041
12. Sex												1.000	-.063	.026
13. Father's occupation													1.000	.160
14. Home town size														1.000

* 5% level of significance = .195.

chance. The apparent conclusion is that the AVL bears little relationship to consumer-choice preferences. It is suggested that there is a need for exploring the possibility of consumer values in a more detailed fashion, perhaps in a manner similar to that used by Shartle (1958) in the context of organizations. Value dimensions developed along these lines appear to have substantial predictability of choice behavior (e.g., see Brumback, 1963). Thus, value dimensions developed specifically with the context of consumer behavior might be expected to considerably improve predictions of consumer choice.

The results of the factor analysis were quite interesting. Certainly it demonstrated that respondents' choices tended to structure themselves along clearly definable dimensions both as a function of product characteristic and the products themselves. Indeed, the first four factors (quality, cost, performance, and style) were able to account for the majority of the total factor variance. Thus, the product is evaluated in terms of the dimensions, as was seen from the significant main effect and interaction in the variance analysis, and these dimensions also emerge as distinct factors in most cases. Similarly, many clearly definable "product" factors emerged in accounting for the remaining variance.

REFERENCES

ALBERT, ETHEL M., & KLUCKHOHN, C. *A selected bibliography on values, ethics, and esthetics in the behavioral sciences and philosophy, 1920-1958*. Glencoe, Ill.: Free Press, 1959.

- BENSON, P. H. A model for the analysis of consumer preference and an exploratory test. *J. appl. Psychol.*, 1955, **39**, 375-381.
- BENSON, P. H., & PERYAM, D. R. Preference for foods in relation to cost. *J. appl. Psychol.*, 1958, **42**, 171-174.
- BROWN, R. L. Wrapper influence on the perception of freshness in bread. *J. appl. Psychol.*, 1958, **42**, 257-260.
- BRUMBACK, G. B. Exploitative game behavior as a function of the individual's exploitative value judgments and of his opponent's strategy and success. Unpublished doctoral dissertation, Ohio State University, 1963.
- DUKES, W. F. Psychological studies of values. *Psychol. Bull.*, 1955, **52**, 24-50.
- EDWARDS, W. The theory of decision making. *Psychol. Bull.*, 1954, **51**, 380-417.
- HAYES, S. P. Some psychological problems of economics. *Psychol. Bull.*, 1950, **47**, 289-330.
- KATONA, G. *Psychological analysis of economic behavior*. New York: McGraw-Hill, 1951.
- KOPONEN, A. Personality characteristics of purchasers. *J. adv. Res.*, 1960, **1**, 6-12.
- MORRIS, C. *Varieties of human value*. Chicago: Univer. Chicago Press, 1956.
- MUELLER, E. Effects of consumer attitudes on purchases. *Amer. Econ. Rev.*, 1957, **47**, 946-965.
- SHEPARD, JANE A., & BAYTON, J. A. *Men's preferences among wool suits, coats, and jackets*. (Bull. No. 64) Washington, D. C.: U. S. Department of Agriculture, Bureau of Agricultural Economics, 1951.
- SHARTLE, C. L. A theoretical framework for the study of behavior in organizations. In A. W. Halpin (Ed.), *Administrative theory in education*. Chicago: University of Chicago, Midwest Administration Center, 1958. Pp. 73-88.
- THURSTONE, L. L. *The measurement of values*. Chicago: Univer. Chicago Press, 1959.
- TUCKER, W. T., & PAINTER, J. J. Personality and product use. *J. appl. Psychol.*, 1961, **45**, 325-329.

(Received July 8, 1963)

TRACKING WITH A DIFFERENTIAL BRIGHTNESS DISPLAY:

II. PERIPHERAL TRACKING¹

STANLEY M. MOSS

*Ohio State University*²

Tracking performances were compared using a differential brightness display (DBD) and a conventional positional display when both displays were moved to the periphery. Results of a previous study and the present study have shown that performance with the positional display was superior to performance with the DBD when the S looked directly at them. As the displays were moved to the peripheral visual areas (15°, 30°, and 45° eccentricity) the reverse of this was true: Performances with the DBD were superior. This was reflected both in the actual performance scores and the amount of time S spent looking toward the displays. These results were interpreted in terms of the underlying physiology of the retina and control movements.

A recent study of tracking (Moss, 1964) had shown that performance with a standard positional display was superior when compared with the performance with a differential brightness display (DBD). Although the magnitude of this difference between performances was significant the percentage reduction of system error was minimal. Considering this final fact it was felt that there may be some aspect of visual tracking where the performance with the DBD may equal if not surpass performance with the conventional positional display.

Recent research (Howell & Briggs, 1959) has indicated the overall performance deterioration of a two-dimensional positional-type control task as one dimension of the task was shifted to the periphery. Performance scores were a decreasing monotonic function of the degree of separation between the two dimensions of the task. Although similar investigations have not been carried out with brightness display, there is a body of research that seems to indicate its merit as a peripheral task display. The use of flashing or varying brightness lights has proved to be helpful in conveying discrete information necessary

in peripheral warning displays (Baker & Grether, 1954). The application of this to a continuous function could provide a highly desirable and unique control system display for a peripheral task.

If the ability to attain skilled performance with a positional-type display can in part be attributed to the sensory function of visual acuity then a brief review of some basic psychophysical and physiological data could support the suggestion made above.

At high intensities of illumination, acuity shows a maximum at the fovea (Mandelbaum & Sloan, 1947). Decrement in acuity increases as the angular separation (in degrees) from the fovea increases: 75 percent at 15°, 92 percent at 32°, and 95 percent at 45°. At low intensities peak visual acuity drops off less gradually as the angular distance from the fovea increases.

The literature concerning differential brightness contrast as a function of angular separation from the fovea is scanty. Steinhardt (1938) states that in the periphery the cones are less sensitive to differential intensities, while the reverse is true for the rods. For visual fields of comparable size and illumination, brightness discrimination ($\Delta I/I$) was .089 at 0° and .224 at 4° eccentricity.

On the physiological side this differentiation can be explained in terms of neural activities underlying each of these processes. Visual acuity depends on the density of receptor cells, while the sensation of brightness is related to the recruitment of the receptor

¹ This study is in part based on a dissertation submitted to the Graduate School of the Ohio State University in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Psychology. The author is indebted to George E. Briggs for advice and assistance throughout the course of this investigation.

² Now at the Mental Health Research Institute, University of Michigan.

cells: the greater intensity of the visual stimulus, the greater the number of rods and cones activated. Hartline (1942) demonstrated this with the excised vertebrate eye. He found convergent pathways to the ganglion cells of the retina which are more prevalent in the peripheral portions than in the fovea. Because spatial summation takes place in the retina, we are able to see dimly illuminated objects. This sensitivity is achieved at the expense of visual resolution; we cannot distinguish fine detail in the periphery because there are many sensory elements that correspond to each ganglion cell in the peripheral retina. This difference in the sensitivities of visual acuity and brightness could have application to the design of certain tasks. Since an individual can perceive shifts in brightness in the periphery more readily than movement in the periphery, the use of a brightness discrimination task may prove to be *more* effective as a peripheral task.

The present study is then concerned with the investigation of problems similar to those stated above in which the tracking of a continuous coherent function was used as the task, using both a positional type and a DBD.

METHOD

Apparatus. A conventional compensatory tracking schema was used with a 12-cycles-per-minute sine wave input. The positional display consisted of a split line generated on the face of an oscilloscope, with the displacement being in the vertical axis. The DBD consisted of a horizontally split visual field, each half consisting of an electroluminescent lamp $1\frac{1}{2} \times 2$ inches with a $\frac{1}{8}$ -inch separation between them. A more detailed description of this apparatus may be found in a recent article (Moss, 1964).

Subjects. Six subjects (Ss) were selected from a previous study (Moss, 1964). They had extensive training with both displays and were selected from a larger population of Ss on the basis of their proficiency with both displays.

Procedure. Each S sat in a chair with his head fixed in a Bausch and Lomb chin and forehead rest 24 inches from a $\frac{1}{4}$ inch red jeweled light that represented the visual fixation point. The displays were placed on a table at various distances to the right of the fixation that corresponded to 15°, 30°, and 45° visual angle from the center of each display. The relative size of each display was kept constant by moving it further away from the front of the table as the visual angle was decreased.

This corresponded to an arc with a radius of approximately 34 inches from S. At all times the frontal

planes of the displays were kept at a 90° angle from the line of sight, and the control stick was placed directly in front of S. The S was instructed to fixate the jeweled light and to track "out of the corner of his eye." He was told that he could shift his eyes over to the display if he felt it were necessary but to avoid this as much as possible. Ambient illumination was raised to .04 foot candles at the S's eyes to enable experimenter (E) to observe eye movements.

One of the six possible sequences of presentation of the three visual angles was randomly assigned, one each, to each of the six S's. The order of display presentation (positional and brightness) was balanced across the six S's. This procedure was repeated, assigning different visual angle presentations to each S and reversing the order of display presentation for each S.

A fourth level of angular separation (0°) was added. Prior to the first sequence, each S tracked as with the other three levels, but his chair and control were rotated so that he fixated directly on the display. The order of display presentation was the same as that for the first sequence. This procedure was repeated at the completion of the last sequence with the order of display presentation being the same as that sequence.

The four separation \times two display \times two order conditions for each S were presented over a period of four days. Each S performed for four blocks of four 65-second trials each. There was a 30-second interval between each trial and a 3-minute rest between blocks. Integrated absolute error scores were recorded for each S at the completion of each trial. These scores were computed during the final 60 seconds of each trial. During this time the E recorded the number of times and amount of time S glanced over to the display by depressing a switch which activated an electric counter and Standard Timer. At the completion of each trial, the S was told his score on that trial.

RESULTS

The results of this study are summarized in Figures 1 through 3. Figure 1 shows tracking performance as a function of peripheral separation from a given fixation point, with displays as the second independent variable. The ordinate is presented in both voltage units (left) and the units of the psychophysical parameters for each display (right). Each point represents the sum of the mean of four trials averaged across the six Ss. Figure 2 shows the average number of glances away from the fixation toward the display as a function of peripheral separation for each display. Figure 3 shows the average time, in seconds, that S glanced away from fixation toward the display as a function of peripheral

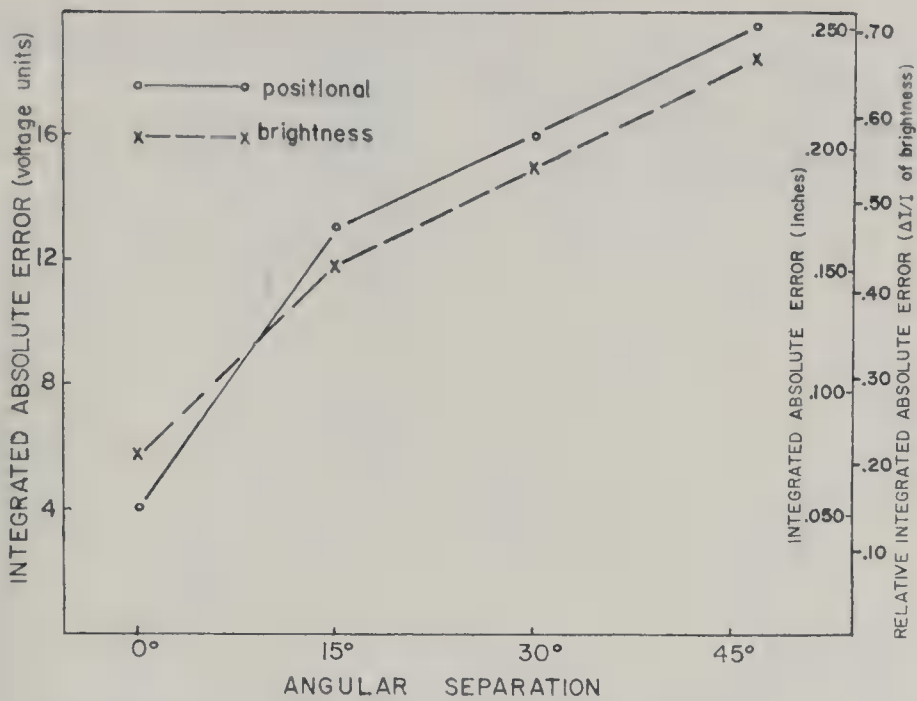


FIG. 1. Tracking performance as a function of visual peripheral separation.

separation for each display. The points in Figures 2 and 3 are the sum of the mean of four trials for each dependent variable averaged across the six Ss. The points for 0° separation were introduced graphically even though S fixated directly on the display during this condition.

Tracking performance, as depicted in Figure 1, reflects a monotonic decrement as peripheral separation increases. Differences between displays are small, showing a cross over between 0° and 15° peripheral separation.

The performance differences between displays remains constant between 15° and 45°. If performance at 0° separation is assumed to be optimal for that display, then the effects of separation from 15° to 45° show a differential effect on performance with each display, as illustrated in Table 1.

Stated in terms of the percentage increase in error from the base error at 0° it becomes apparent that larger magnitudes of peripheral separation reflect greater performance differences between displays. The rate of change

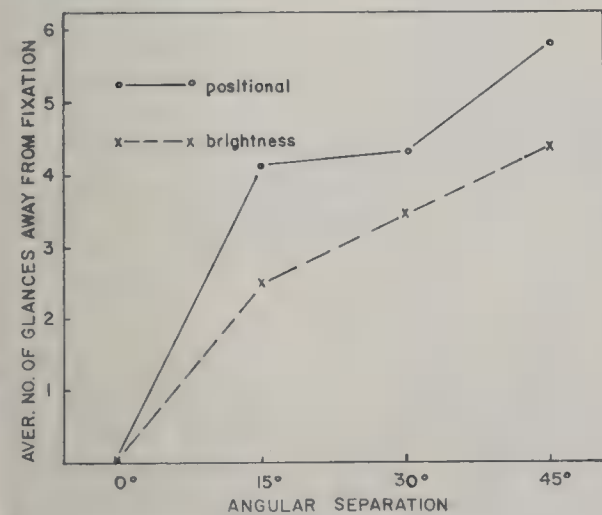


FIG. 2. Average number of glances away from fixation as a function of peripheral separation.

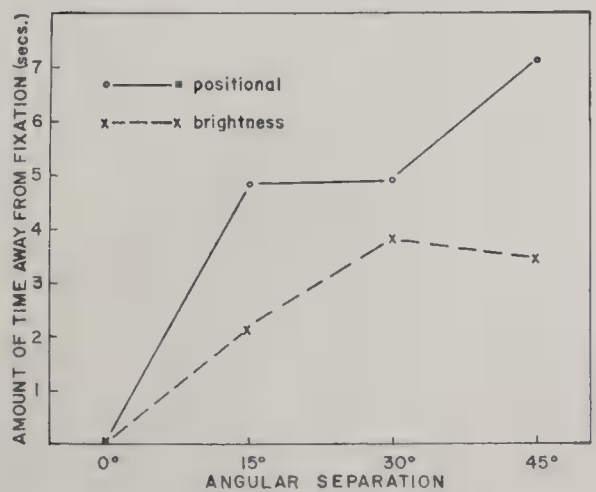


FIG. 3. Average time away from fixation as a function of peripheral separation.

TABLE 1
PERCENTAGE INCREASE IN TRACKING ERROR FROM
THAT AT 0° PERIPHERAL SEPARATION

Separation	Positional	Brightness	Difference
15°	325	205	120
30°	400	263	137
45°	488	340	148

of these differences does not increase as predicted earlier.

Since the present study is concerned with peripheral tracking, i.e., tracking out of the corner of the eye, larger values in Figures 2 and 3 are meaningful only if they are viewed as being detrimental to the task. Viewed in this manner, these graphs reflect results similar to those shown in the graph of Figure 1. Figure 2 shows an increase in the number of glances toward the display as peripheral separation increases. Throughout all levels of peripheral separation, *S* glances toward the positional display more than he does toward the brightness display. Similar results may be observed in Figure 3: *S* spends more time looking toward the positional display than toward the brightness display through all levels of separation. The points representing the positional display show the same characteristics as their counterparts in Figure 2, while the points depicting performance with the brightness display show a decrease in the amount of time *S* spent looking from 30° to 45° as compared with an increase between these two points in Figure 2. When the results of Figure 2 and 3 are combined, this disparity is seen as a decrease in the average amount of time per glance at 45° separation with the brightness display, 1.10 seconds at 30° and .80 seconds at 45°. The decrease in time at this point could explain the difference in the rate of change of the differences between displays in Table 1: *S* spent comparatively less time looking at the brightness display at this separation than at the separation of 30°.

Analyses of variance were applied to the data summarized in Figures 1 through 3. In all cases the main effects and their interactions were tested against their higher-order

interactions with *Ss*. The Order effects were introduced only to account for this source of variance.

In the analysis of the absolute error scores no significance arose between the display groups. This can be explained by the cross over between 0° and 15° and by the relatively small differences between displays at all levels of angular separation. Angular separation was significant at $p < .005$ with the means in the predicted direction. The Duncan Multiple Range Test (Edwards, 1960) was applied to the means of the data in the significant $D \times A$ interaction ($.025 < p < .05$). The only significance ($p = .05$) between displays at the same angular separations arises at 0° and 30°. The lack of significance at 45° can be explained by the fact that *S* spent less time glancing over to the brightness display at this point, as stated above. A closer look at Figure 3 will indicate that *S* spent almost as much time looking toward the positional display at 15° (4.8 seconds) separation as he did at 30° (4.9 seconds) separation, whereas *S* spent 2.2 seconds at 15°, and 3.8 seconds at 30° using the brightness display. These differences help explain the lack of significance between displays at 15°.

The results of the analyses of the number of glances and time scores reflect significant differences between the displays at $.025 < p < .05$. The differences between angular separations are significant at $.025 < p < .05$ for the number of glances but show a lack of significance for the time scores. Again, this lack of significance can be attributed to *S*'s performance with the brightness display at 45°. It should be stated that the analysis was computed only at the 15°, 30°, and 45° points thus making the differences between separations appear more obvious in Figures 2 and 3 because of the included 0° separation point.

DISCUSSION

Even though the present task was of a less complex nature, the above results are in agreement with those obtained by Howell and Briggs, which indicated a monotonic decrement in tracking performance as angular separation is increased. The differential effects of

displays on performance as angular separation was increased are in the direction predicted. The magnitude of these differences is obscured by the experimental artifact of allowing *S* to glance over to the display. It is felt that were *S* allowed only to fixate and track out the corner of his eye, the magnitude of these differences would have reflected significant differences between display performances at the three levels of angular separation.

These results are in accord with the predictions based on the psychophysical data mentioned earlier. It would be misleading to make direct comparisons between the psychophysical equivalents of tracking performance and their counterparts in a well-controlled psychophysical study, basically because of the artifact mentioned above. Discussion then will be limited to the relative deterioration of tracking performance induced by peripheral separation as related to the underlying retinal changes. In this case we are dealing only with visually encoded information and not with the control components of performance.

As stated earlier, visual resolution in the periphery is decreased because of the multiple convergence of the receptor cells to the ganglion cells of the retina. When one object increases in luminosity, spatial summation takes place, allowing more ganglion cells to fire and hence to induce the perception of dimly illuminated objects. The closer to the fovea, the less the convergence and the greater the increase in resolution. (We are dealing with intensities that are well into the scotopic levels of sensitivity.) This variation in structure, while maintaining spatial summation effects across the retina, appears as a monotonic decrement in visual acuity when test objects are moved to the periphery. If we take these facts into consideration, the differential decrements between displays at larger peripheral separations are understandable. Larger degrees of angular separation reduced the resolution between the target and cursor of the positional display, thus conveying less information to *S* regarding his continuous proficiency level. Larger degrees of angular separation with the brightness display, while maintaining spatial summation effects, conveyed more information than the positional

display. Support is given to this notion by *S*'s statements, at the completion of the testing session, that he could tell when the two squares of the brightness display were of comparable brightness but found it more difficult to tell when the two lines of the positional display were near each other.

The differences between displays at increasing degrees of angular separation can also be explained in terms of the conclusions drawn from an earlier study (Moss, 1964). It was stated that while performing with the positional display *S* makes more small adjusting control movements by trying to keep the cursor aligned with the target. The *S* makes smoother control movements while performing with the brightness display because of an "area of uncertainty" around zero error. When separations are introduced between fixation and the display, *S* persists in making smoother control movements with the brightness display while relying less on visually presented information; with the positional display he relies more on the visually presented information and therefore makes similar adjustive movements. Apparently, *S* needs to obtain more visual information with the positional display. The data summarized in Figures 2 and 3 support this notion. Throughout all levels of angular separation, *S* spends more time looking at the positional display than at the brightness display; thus he obtains more information regarding the magnitude of performance proficiency while tracking.

There is a relatively short amount of time during each trial that *S* spends looking toward the display. At 45° separation, *S* spends 12 percent of the total trial time looking toward the positional display and only 6 percent of the total trial time looking toward the brightness display. It is reasonable to assume from this that *S* is making continuous control movements with both displays and only looks toward the display when the magnitude of the error becomes large enough to be perceived in the periphery. With the positional display, this information is so degraded that *S*'s tendency to glance toward the display is increased in order to maintain an acceptable performance level.

REFERENCES

- BAKER, C. A., & GREYER, W. F. Visual presentation of information. *USAF WADC tech. Rep.*, 1954, No. 54-160.
- EDWARDS, A. L. *Experimental design in psychological research*. New York: Rinehart, 1960.
- HARTLINE, H. K. The neural mechanisms of vision. In, *The Harvey Lectures*, 1941-42. Series XXXVII. Lancaster, Pa.: The Science Press Printing Company, 1942, 39-68.
- HOWELL, W. C., & BRIGGS, G. E. The relative importance of time sharing at central and peripheral levels. *Technical Report: NAVTRA-DEVCEN* 508-2, October, 1959b, Port Washington, New York.
- MANDELBAUM, J., & SLOAN, L. L. Peripheral visual acuity: With special reference to scotopic illumination. *Amer. J. Ophthal.*, 1947, 30, 581-588.
- MOSS, S. M. Tracking with a differential brightness display: I. Acquisition and transfer. *J. appl. Psychol.*, 1964, 48, 115-122.
- STEINHARDT, J. Intensity discrimination in the human eye, I. The relation of I/I to intensity. *J. gen. Physiol.*, 1938, 20, 185-209.

(Received July 17, 1963)

THE EFFECT OF TIME STRESS AND THE ELIMINATION OF CUE INFORMATION ON THE DISPLAY-CONTROL RELATIONSHIPS OF MOVING SCALE INSTRUMENTS

A. B. HILL AND C. Q. LARGE

Army Operational Research Establishment, Farnborough, England

The reading of 4 rotary-moving scale assemblies of different display-control relationships was investigated using 16 Ss, 8 soldiers and 8 scientists, in a Latin-square design. Performance times, initial movement errors, and final setting errors were recorded for runs of 60 dial settings on each assembly. Initial movement errors indicated that an assembly employing compatible movement, scale numbers increasing from left to right, and a "turn anti-clockwise to increase" condition was optimum. Performance times tended to reflect the number of errors committed. No final setting errors were made, this being attributed to the nongraduated type of scale used. The absence of cue information provided by the visibility of 2 or more scale values was found to be detrimental to performance.

The display-control relationships for moving scale instruments pose several interesting problems. Their use presents what is essentially a three-variable situation in that: the control knob may be turned clockwise or anti-clockwise to register a different scale setting, the scale may be made to move either clockwise or anticlockwise, and the scale values may be made to increase from either left to right or right to left.

It is usually advocated that rotary controls should be turned clockwise to increase the setting, that scale values should increase from left to right, and that the scale itself should move in the same direction as the rotary control. Bradley, (1954) recognizing that one of these conventions must be broken in a normal moving scale assembly showed that the "clockwise rotation of the control to increase the setting" is the principle which may be most successfully ignored. In the case being reported an experiment similar to number one in Bradley's series was undertaken, but the subjects (Ss) were exposed to subjective time stressing. This time stress was imposed by instructing Ss to work as quickly as possible and allowing them to see that they were being timed. This procedure was considered to be closer to the real life situation than allowing a set time for the task. Another important difference lay in the type of scale used, Bradley's containing 100 graduations; every fifth graduation mark being longer and thicker

than the "unit" marks and every tenth mark being longer and thicker still, and numbered. Two "ten" marks were always visible through the window thus providing cue information as to which way the *dial* should move. In the present case a discrete digit scale, without graduations, was used with only a single digit visible, thus eliminating cue information.

It has been shown that for rotary controls a stereotype "turn clockwise to increase" exists but is claimed to take effect only when such an increase does not cause movement in an observed display (Bradley, 1959). This claim is supported by his findings on moving scale instruments (Bradley, 1954). The purpose of the present experiment was to determine whether these findings for moving scale instruments would break down and be replaced by population stereotype behavior under conditions of time stress and no cue information.

METHOD

The apparatus consisted of a box, one face of which was used as a vertical display panel. Along the top and bottom edges of this panel six windows $\frac{3}{8} \times \frac{1}{16}$ inch were cut with a spacing of $2\frac{1}{4}$ inches between their centers. Rotary knobs 1 inch in diameter were mounted on the panel each being spaced 3 inches vertically above or below a window, thus forming two rows of six knobs per row. The internal mechanism of the box was so arranged that a change from direct to indirect drive from control to display could be obtained by turning the box upside down. Twelve dials each of $2\frac{1}{2}$ inches in diameter were arranged so as to move behind the 12

windows. The dials carried black figures $\frac{5}{8}$ inch high on a white background with a character height/width ratio of 2:1 and character stroke/height ratio of 1:5. Six of the dials were arranged in a row so that the numerals increased from left to right, (0-9), while the numerals on the other six increased from right to left, (0-9).

The apparatus was capable of supplying four experimental conditions.

- A—Direct drive (i.e., the scale moved in the same direction as the control knob).
Scale numbers increasing from left to right.
Turn anticlockwise to increase the setting.
- B—Indirect drive (i.e., the scale moved in the opposite direction to the control knob).
Scale numbers increasing from right to left.
Turn anticlockwise to increase the setting.
- C—Direct drive
Scale numbers increasing from right to left.
Turn clockwise to increase the setting.
- D—Indirect drive
Scale numbers increasing from left to right.
Turn clockwise to increase the setting.

Sixteen Ss were used, eight were members of the laboratory staff and eight soldiers. The S was seated with the apparatus in front of him, the windows at approximately eye level. Before each run S was allowed to familiarize himself with the experimental condition under test. When S reported that he was ready, the experimenter set all six dials to "five" and placed before S a file card on which was clearly printed a six figure random number. The S's task was to set each figure of the number into the appropriate window and having completed this to reset the dials to read five.

In each run the S was presented with five such file cards in random order, thus producing a total of 60 settings. He was timed from the presentation of the first card to the completion of the resetting task on the fifth card.

A Latin-square design was used to randomize the order of experimental conditions.

The instructions given to Ss before each run were:

You will be presented with a card on which is printed a six figure number. Your task is to set

this number into the machine so that it appears in the windows above the control knobs. You will see that at first each window shows the figure "5." When you have set in the number on the card you are required to reset the machine so that all the windows show the figure 5 again, when you will be given another card.

There are five numbered cards in all. You will be timed in this task and it is important that you do it as *quickly and accurately as you can*.

You will be allowed some practice to get used to the apparatus.

If there is anything you do not understand please ask the Experimenters.

Initial direction of movement errors for each control knob and final setting errors were recorded along with the time for each run. The Ss were interviewed informally after the experiment.

RESULTS

Table 1 shows the time in seconds taken for one run on each of the four conditions by the two groups of Ss.

A Friedman Two-Way Analysis of Variance carried out on the times for both groups of Ss showed a significant difference over the four experimental conditions for the scientists ($p < .05$) but no significant difference for the soldiers. The soldiers tended to perform slightly more quickly than the scientists, but this tendency was not statistically significant. The implications of these results will be discussed later when they will be related to initial movement error patterns.

The initial movement errors are shown in Table 2.

Analysis of the scientists' results in Table 2 by means of a chi-square test (using Yate's correction for continuity) showed a significant difference between Condition A and each of the other three conditions, ($p < .001$). No

TABLE 1
TIME IN SECONDS FOR ONE RUN ON EACH EXPERIMENTAL CONDITION

Scientists Conditions					Soldiers Conditions				
Subjects	A	B	C	D	Subjects	A	B	C	D
1	116	127	123	123	1	103	98	95	130
2	173	180	175	197	2	122	151	149	125
3	130	160	143	157	3	105	110	107	105
4	103	124	112	123	4	105	127	80	135
5	158	170	166	165	5	135	155	140	128
6	217	142	144	190	6	100	145	110	120
7	123	135	127	156	7	165	156	123	170
8	123	170	174	176	8	120	145	135	113
Total	1143	1208	1164	1281	Total	955	1087	939	1026

TABLE 2
INITIAL MOVEMENT ERRORS FOR EACH OF THE EXPERIMENTAL CONDITIONS

Subjects	Scientists Conditions				Subjects	Soldiers Conditions			
	A	B	C	D		A	B	C	D
1	2	13	14	6	1	2	2	1	12
2	2	14	20	12	2	4	6	8	10
3	4	25	5	5	3	9	3	11	6
4	4	9	12	6	4	4	12	4	13
5	1	19	15	20	5	7	5	4	8
6	32	2	0	10	6	1	12	4	8
7	0	9	8	5	7	10	12	13	18
8	4	18	22	22	8	8	13	10	6
Total	49	109	96	86	Total	45	65	45	82

significant differences were found between any two of the other three conditions at the 5% level using the same test.

The soldiers' results in Table 2 show overall lower levels of initial movement errors in comparison to those of the scientists. Chi-square tests (again using Yate's correction for continuity) showed no significant differences between Conditions A, B, and C at the 5% level, nor between Conditions B and D although there was a highly significant difference ($p < .001$) between Condition A or C and Condition D.

The error patterns for the two groups of Ss were shown by a chi-square test to be significantly different ($p < .01$).

It is clear that the initial-error scores for the scientists are decisive with Condition A being markedly better than the other three which are not significantly different from one another.

The initial movement error pattern for the soldiers is not so clear-cut and at first sight appears confusing. In this case no significant differences were found between A, B, and C, although Condition B was found to be on the border line of significance when compared with A and C. The explanation for this is one of training, as the soldiers had previously been trained on a piece of equipment embodying an assembly of Type B.

No final setting errors were committed in either group of Ss. This finding is comparable to Bradley's which showed few errors of this type. The differences such as they were can be attributed to the types of scales used.

On interview several Ss mentioned that with a normally numbered dial (e.g., a speed-

ometer) the digits increase in a clockwise direction, and that therefore if the dial must be turned rather than a pointer, the dial must move anticlockwise to show a higher number.

DISCUSSION

A appears in both groups to be a good assembly, as judged from initial movement error scores. For the group of soldiers it must be admitted that A was not found to be significantly different from B or C, but in this case one must bear in mind that these Ss had been trained on equipment which embodied an assembly of Type B.

The relation of times to error scores is one in which the group which produced the quicker and less variable times, the soldiers, also produced the lower error scores. This is an important finding since it shows that time of operation was not the most important variable in the task, i.e., errors were mainly produced by the inherent difficulty of the assemblies themselves. This is not to say that time did not stress the Ss and contribute towards errors but rather that it was of secondary importance to the design of the assembly. The variation in performance times over the four experimental conditions shown by the scientists is best explained as being a function of error score, and hence of assembly difficulty. This argument is supported by the performance times and error scores for the soldiers which showed no significant variation over the four conditions for times, and no significant difference in errors between three of the four conditions. As Condition A was executed most quickly by the scientists, and as no significant differences were found

in times over the four conditions for the soldiers, it is reasonable to suppose that there is at least a tendency for assemblies of Type A to be operated more quickly as well as more accurately than those of other types.

In Bradley's first experiment (1954) Ss were given cue information by the labelled "ten marks," two of which were always visible. In the present experiment no such cue information was available as the window behind which the dial moved was capable of showing only one digit at any one time. From the absence of final setting errors in the work being reported here, it would be tempting to conclude that the elimination of the kind of cue information available to Bradley's Ss improves accuracy of performance. There is, however, another factor which may have contributed to the absence of final setting errors, and that is the type of scale used. The graduated scale used by Bradley required interpolation on the part of the S, whereas the one used by the authors did not, and it seems likely that his final setting errors were in fact errors of interpolation. This must be so as it is difficult to see why there should be a causal link between the elimination of cue information and the elimination of final setting errors.

It would be reasonable to hypothesize that the elimination of cue information would tend to produce more initial errors. A percentage comparison with Bradley's results showed that his initial movement error scores were considerably lower than those found by the authors for Conditions A and C. However the reverse was true for Conditions B and D. The implication here is that for assemblies in which dial movement and control knob movement are compatible, eliminating cue information increases initial movement errors. The reason for this is that when the S can see which way the dial must move to register a desired setting he tends to expect compatibility between control and display (a strong preference exists for compatibility—McCormick, 1957), and thus he turns the knob in the direction he wants the dial to move. For assemblies which do in fact have compatible display-control relationships this strategy is appropriate. For assemblies which are not compatible the strategy is inappropri-

ate and results in increased initial movement errors.

Eliminating the cue information means that the S has no *visual* evidence concerning the direction in which the dial must move to register the desired setting and thus he does not readily assume compatibility. One therefore expects that with a noncompatible assembly S will perform at least as well, if not better, in the absence of visual cues. Without cues he is not being urged by visual evidence to adopt an incorrect strategy.

This elimination of cue information does not seriously affect *relative performance* over the various types of assembly. That is, compatible assemblies are still *relatively* better than the noncompatible.

A possible reason for the supremacy of Type A assembly over Type C was suggested by the interviews. In the absence of cue information some of the Ss claimed to visualize a "normal" dial with numbers increasing clockwise. They realized that it had to be turned anticlockwise to increase the setting and expected a compatible relation between control and display. Although such an explanation was not elicited from all the Ss it is possible that this process may operate at an unconscious level.

There are four conclusions to this experiment which would be of use to the designer of moving scale instruments. These are:

1. Display-control movements should be compatible.
2. The control should turn anticlockwise to increase the setting.
3. A nongraduated scale should be used whenever possible.
4. Cue information should be provided by the use of a window large enough to show more than one scale value at any given time.

REFERENCES

- BRADLEY, J. V. Desirable control-display relationships for moving scale instruments. USAF WADC *tech. Rep.*, 1954, No. 54-423.
- BRADLEY, J. V. Direction-of-knob-turn stereotypes. *J. appl. Psychol.*, 1959, 43, 21-24.
- MCCORMICK, E. J. *Human engineering*. New York: McGraw-Hill, 1957.
- SIEGEL, S. *Nonparametric statistics*. New York: McGraw-Hill, 1957.

(Received August 2, 1963)

SUPERVISOR PERCEPTION OF WORK GROUP MORALE¹

THOMAS H. JERDEE

School of Business Administration, University of North Carolina

Objectives were to determine the degree to which an "accuracy" measure of supervisory empathy is influenced by unrealistic estimation tendencies common among factory supervisors and to determine whether "group" empathy exists among supervisors. Subjects (Ss) were 38 supervisors and 190 subordinates in 3 plants. Measures were group morale (GM), predicted group morale (PGM), accuracy (ACC), typicality of prediction (TYP), and general effectiveness. ACC was negatively related to PGM ($r = -.38, p < .05$) and positively related to TYP ($r = .43, p < .01$). The correlation between GM and PGM was $-.05$. Conclusions are that "accuracy" measures of supervisory empathy are of questionable value and that supervisors in the 3 plants were not able to react differentially to group morale.

Accurate perception of subordinate attitudes frequently has been suggested as an important ability in factory supervision (Bel lows, Gilson, & Odiorne, 1962, pp. 63-65; Tannenbaum, Weschler, & Massarik, 1961, pp. 37, 43-44). This ability, often called "supervisory empathy," has been studied by several investigators, and their results, if taken at face value, indicate that empathy may be an important supervisory trait (Browne & Shore, 1956; R. J. Johnson, 1954; Jones, 1954; Nagle, 1954; Patton, 1954). However, as Gage and Cronbach (1955) have pointed out, many of these studies have been subject to some rather serious methodological limitations, and there is a need for further research on supervisory empathy.

PROBLEM

The present study was concerned with supervisory awareness of the level of work group morale, a type of empathy that may be important even in situations where supervisors have few direct contacts with individual subordinates. This type of empathy customarily is measured by comparing supervisors' predictions with subordinates' responses to morale-scale items. A supervisor is considered skillful in group empathy if he can predict accurately the level of morale among his subordinates.

One objective of the present study was to determine the degree to which this customary "accuracy" measure of supervisory empathy is influenced by a general tendency of supervisors to overestimate or to underestimate subordinate morale, or to vary their estimates too much or too little. As Gage and Cronbach (1955) have pointed out, these tendencies would make the interpretation of accuracy measures difficult.

Suppose, for example, that supervisors generally tend to make uniformly high estimates of subordinate morale. If their subordinate groups actually differ in morale, supervisors of high-morale groups will get favorable accuracy scores and supervisors of low-morale groups will get unfavorable accuracy scores. The result will be a positive relation between supervisory accuracy and observed morale, and it might be tempting to conclude that the more accurate supervisors have used their empathic skill to achieve higher subordinate morale. But if supervisors generally tend to make uniformly high estimates of subordinate morale, the greater accuracy of some supervisors may simply reflect their good fortune in having groups with higher morale. As Gage and Cronbach point out, the more parsimonious interpretation of a situation like this is that the accuracy measure reflects luck rather than empathy. There is no basis for making inferences about supervisory differences in empathic ability. In fact, the differences in supervisory accuracy do not even prove that empathy exists as a variable supervisory trait.

¹ This study was based on data collected for the author's doctoral dissertation, completed under Marvin D. Dunnette at the University of Minnesota. Appreciation is extended to Carroll Stein, who helped greatly in collecting the data.

In the example just discussed, interpretation of the accuracy measure is difficult because all supervisors have overestimated the level of subordinate morale, and each supervisor's accuracy may depend more on the level of his subordinates' morale than on his own empathic skill. From a logical standpoint it is clear that any general tendency of supervisors to overestimate or underestimate subordinate morale, or to vary their estimates too much or too little, could cause difficulties in the interpretation of accuracy measures. The actual nature and practical importance of these difficulties were investigated in the present research.

Another objective of the present study was to determine whether "group" empathy exists at all, i.e., whether factory supervisors are able to react differentially to the morale of their own subordinate work groups. In spite of the methodological problems previously discussed in connection with customary accuracy measures, it is possible that data in the form of predicted and observed morale may contribute useful information about supervisory empathy, if they are subjected to correlational analysis. A positive relation between predicted group morale and observed group morale would indicate the influence of factors other than meaningless response tendencies, even though it would not provide a measure of individual supervisory empathy. This correlational approach to the exploration of empathy was applied in the present research. The correlation between supervisory perceptions of group morale and actual group morale was examined.

METHOD

Subjects

Subjects were 38 first-line factory supervisors and 190 of their subordinates in three unionized manufacturing plants. All three plants manufacture machinery and metalware requiring a variety of metalworking and assembly operations.

Measures

The basic measuring instrument was a Likert-type morale scale containing 20 items from the Triple Audit Employee Attitude Scale developed by the Industrial Relations Center of the University of Minnesota. Total score on these 20 items correlates approximately $+0.96$ with total score on the complete

54-item scale, which has an odd-even reliability of $+0.93$ (Yoder, Heneman, & Fox, 1954). The morale scale was completed by a random sample of five subordinates for each of the 38 participating supervisors, under conditions of assured anonymity.

The measure of *group morale* for each supervisor was the average 20-item score for his five subordinates. The measure of *predicted group morale* for each supervisor was the average 20-item score based on his predictions of his subordinates' responses. The *accuracy* measure for each supervisor was the difference between group morale and predicted group morale (a smaller difference resulted in a higher accuracy score). The *typicality* of prediction for each supervisor was the difference between his predicted group morale and the mean predicted group morale for the supervisors in his plant (a smaller difference resulted in a higher typicality score). As a matter of additional interest, each supervisor's *general effectiveness* as a supervisor was measured by rankings made by his superiors. In one plant the rankings were made only by the Plant Manager, and no information was available regarding their reliability. In the second plant the rankings were made by four persons in high-level managerial positions. One person's rankings were eliminated because of their low correlation with the other three persons' rankings. The Horst reliability coefficient for the remaining three rankers was $+0.92$ (this was a biased coefficient, based on only three of the original four raters) (Horst, 1949). The median rank of these three was used as the measure of effectiveness in the second plant. In the third plant, rankings were made by three persons, and the Horst reliability coefficient was $+0.97$. The median rank was used as the measure of effectiveness.

Analysis

Methodological difficulties with the customary distance measure of accuracy were studied in two ways. First, the difference between mean group morale and mean predicted group morale was studied in a two-way analysis of variance (group morale and predicted group morale by plants). If this difference were large, there might be more likelihood of a spurious relation between accuracy and the level of group morale. On the other hand, if there were only a small difference between mean group morale and mean predicted group morale, the typicality of a supervisor's estimate of group morale might be expected to play a more important part in determining his accuracy (unless his variant predicted group morale were based on genuine empathic skill).

The next step was to determine the correlations (in each plant) between accuracy scores and (a) the level of group morale, (b) the level of predicted group morale, and (c) the typicality of predictions. A significant correlation between accuracy scores and any of these three measures would support Gage and Cronbach's (1955) criticisms of accuracy scores as measures of supervisory empathy.

Another objective of this study was to determine whether group empathy exists at all, i.e., whether

factory supervisors are able to react differentially to the morale of their subordinate work groups. This ability was explored by determining the correlation between group morale and predicted morale in each of the three plants. A high positive correlation between group morale and predicted group morale would be an indication that supervisors really are able to respond differentially to the morale of their work groups. A low correlation would indicate the absence of "group" empathy.

RESULTS

A first step in the study of the accuracy measure of empathy was an analysis of the difference between mean group morale and mean predicted group morale. The mean differences for the three plants are shown in Table 1. In all three plants the supervisors predicted higher morale than was observed; but the differences between predicted morale and actual morale were quite small in two

TABLE 1

PREDICTED VERSUS OBSERVED MORALE SCORES

Measure	Plant A	Plant B	Plant C
Number of supervisors	12	13	13
Mean predicted morale	48.5	52.3	52.7
Mean observed morale	46.8	50.5	48.0
Mean difference	1.7	1.8	4.7
Variance of predicted group morale	7.33	4.47	10.51
Variance of observed group morale	4.83	3.45	5.20

of the three plants, and as the analysis of variance in Table 2 shows, the "predictions vs. observations" effect was not significant. Thus the general importance of a tendency to predict unrealistically high morale is not clearly established by this analysis of mean differences. However, further evidence regarding its importance is presented in subsequent paragraphs dealing with the correlations between accuracy scores and other measures.

Although the general level of supervisors' predictions was fairly accurate, the variance of predicted group morale (shown in Table 1) was almost double the variance of observed group morale. In view of this excessive variance of the supervisors' predictions, the typicality of a supervisor's predicted group morale might be expected to play an important part in determining his accuracy. This expectation was borne out in further analysis

TABLE 2

ANALYSIS OF VARIANCE OF MORALE SCORES

Source of variation	df	MS	F	F.95
Plants	2	86.4	2.10	3.13
Predictions versus observations	1	139.6	3.39	3.98
Interaction	2	18.0	.44	3.13
Error	70	41.2		

of the correlations between accuracy scores and other measures.

Correlations between accuracy scores and other measures are shown in Table 3. The correlations in the three plants were combined using Fisher's z transformation, and the combined correlation coefficients were tested for significance (P. O. Johnson, 1949, p. 53). The results lend further support to the importance of the effects of response tendencies on accuracy scores. The combined correlation between supervisors' predicted group morale and the accuracy of their predictions was $-.38$, significant at the .03 level. This is not surprising in view of the tendency of supervisors to predict higher morale than was observed. The higher a supervisor's prediction of group morale, the further he was in error.

The combined correlation between the typicality of supervisors' predictions and their accuracy was $+.43$, significant at the .01 level. This, too, is not surprising, in view of the excessive variability of supervisors' predictions. The supervisors who strayed least from the mean of all supervisors' predictions turned out to be more accurate. The supervisors whose predictions were unusually high or unusually low turned out to be less accurate.

If the supervisors were reacting differen-

TABLE 3

CORRELATIONS BETWEEN VARIABLES

Variable	2	3	4	5
1. Accuracy of prediction	.18	-.38*	.43**	.16
2. Observed morale		-.05	.17	.13
3. Predicted morale			.03	-.12
4. Typicality of prediction				-.03
5. Supervisory effectiveness				

Note.— $N = 38$.

* $p \leq .05$.

** $p \leq .01$.

tially to the morale of their own work groups, there should have been a positive correlation between group morale and predicted group morale. The obtained correlation, combined for the three plants, was $-.05$. In one plant the correlation reached $-.57$, indicating that in this plant the supervisors who thought they had higher work-group morale actually had lower work-group morale. These results indicate that the supervisors in these three plants were not able to react differentially to the morale of their own work groups.

DISCUSSION

Results of this study indicate that the conventional accuracy measures of supervisory empathy are of questionable value, because they are too sensitive to response tendencies that may have nothing to do with genuine empathic insight. In general, supervisors whose predictions are close to the average of all supervisors' predictions, or perhaps slightly on the low side, tend to get higher accuracy scores. But the low correlation between predicted and observed morale indicates that supervisors are not able to react differentially to the morale of their subordinates as a group.

Although these findings do not support the viewpoint that empathy is an important variable in factory supervision, they must be interpreted with caution. This study has examined only the ability of supervisors to perceive the level of work group morale. Even though supervisors do not seem to possess this ability to any great extent, they may still be sensitive to differences in attitudes between individual subordinates. Hatch (1962), using an ingenious forced-choice technique and 73 items constructed especially for his subjects, did demonstrate that this latter type of "individual" empathy exists among sales supervisors. It is possible that results similar to Hatch's might be obtained for factory supervisors, using a comprehensive inventory covering attitudes toward specific factors in the immediate work environment.

In fact, it seems rather obvious that dif-

ferential perception of subordinate characteristics exists and can be measured, providing the appropriate kinds of subordinate characteristics are examined. For example, Tannenbaum, Weschler, & Massarik (1961, p. 38), in their discussion of "levels of depth" in social sensitivity, point out that the average leader might find it easier to estimate a follower's age than to estimate his "unconscious dynamics." The practical question, however, is whether supervisors are able to perceive some of the less obvious but presumably important differences in subordinate likes and dislikes, and in the case of factory supervisors this ability has not been demonstrated.

REFERENCES

- BELLOWS, R. M., GILSON, T. Q., & ODIORNE, G. S. *Executive skills—their dynamics and development*. Englewood Cliffs, N. J.: Prentice-Hall, 1962.
- BROWNE, C. G., & SHORE, R. P. Leadership and predictive abstracting. *J. appl. Psychol.*, 1956, **40**, 112–116.
- GAGE, N. L., & CRONBACH, L. J. Conceptual and methodological problems in interpersonal perception. *Psychol. Rev.*, 1955, **62**, 411–422.
- HATCH, R. S. *An evaluation of a forced-choice differential accuracy approach to the measurement of supervisory empathy*. Englewood Cliffs, N. J.: Prentice-Hall, 1962.
- HORST, P. A. A generalized expression for the reliability of measures. *Psychometrika*, 1949, **14**, 21–31.
- JOHNSON, P. O. *Statistical methods in research*. New York: Prentice-Hall, 1949.
- JOHNSON, R. J. Relationship of employee morale to ability to predict responses. *J. appl. Psychol.*, 1954, **38**, 320–323.
- JONES, D. R. Measuring supervisor awareness of employee attitudes: A comparison of methods. Unpublished doctoral dissertation, Purdue University, 1954. (*Dissert. Abstr.*, 14:2120).
- NAGLE, B. F. Productivity, employee attitudes and supervisory sensitivity. *Personnel Psychol.*, 1954, **7**, 219–234.
- PATTON, W. M., JR. Studies in industrial empathy III. *J. appl. Psychol.*, 1954, **38**, 285–288.
- TANNENBAUM, R., WESCHLER, I. R., & MASSARIK, F. *Leadership and organization: A behavioral science approach*. New York: McGraw-Hill, 1961.
- YODER, D., HENEMAN, H. G., JR., & FOX, H. Auditing your manpower management. Bulletin No. 13. Minneapolis: Industrial Relations Center, University of Minnesota, 1954.

(Received August 8, 1963)

KNOWLEDGE OF PERFORMANCE AS AN INCENTIVE IN REPETITIVE, MONOTONOUS TASKS¹

ALPHONSE CHAPANIS

Johns Hopkins University

This experiment tried to isolate the purely motivational effect of knowledge of performance from its informational and rewarding aspects. The subjects (Ss) worked an hour a day, for 24 days, punching random digits into a teletype tape. They were told they were programming a computer and efforts were made to lend credibility to this fiction. 16 male undergraduate students were assigned to 1 of 4 groups. Ss in 1 group received no information about their output. In 2 other groups, the Ss could see a counter which tallied every stroke on the perforator and could, if they so chose, determine their daily work output. In the 4th group Ss were required to write down their output at the end of 15, 30, and 45 minutes, "for accounting purposes only." No significant differences were discovered among the 4 groups. Suggestions are offered to account for the discrepancy between the results of this experiment and those of a similar experiment reported by Gibbs and Brown in 1955.

That knowledge of results improves performance is perhaps one of the most dependable and thoroughly-tested principles in modern-day psychology (See Ammons, 1954, for example). The principle holds for animals and men, for children and adults, for groups as well as individuals, and for a wide variety of learning, psychomotor, monitoring, and other general performance tasks.

When knowledge of results is manipulated as an experimental variable, the conditions of the experiment are generally arranged in such a way that the subjects (Ss) clearly perceive the information they receive as an integral part of the experimental situation. Under these circumstances, knowledge of performance may serve three functions according to Ammons: it may *inform* S about what he should, or should not, be doing, it may *reward* S for acceptable performance and/or punish him for unacceptable performance, or it may *motivate* S. In 1955, Gibbs and Brown tried to isolate and measure the purely motivational aspects of knowledge of performance by designing an experiment in which knowledge of results was more casual and incidental than

is usually the case. Specifically, they set Ss to work at the uninteresting, unskilled and repetitive task of copying pages from difficult scientific reports, encyclopedias, and historical reviews with a document copying machine. In half of the trials Ss could see a counter which tallied each page as it was copied; in the other half of the trials the counter was covered. The Ss were not instructed to meet daily quotas, they were paid uniformly, they were never supervised nor monitored during their work sessions, and the experimenter (*E*) never made comments about the daily production figures. In short, *Es* tried to arrange conditions in such a way that a *S*'s output would be determined entirely by self-competition and his own satisfaction in working. The results show that when the men could see the counter their output was significantly higher than when they could not. The increase was approximately equal to one extra day's production in every four.

If this finding by Gibbs and Brown (1955) can be shown to hold generally, it would have immensely important practical consequences, and the authors themselves have much to say about the application of their results to a variety of industrial situations. Still, for all of its inherent interest, this original finding does not appear to have been duplicated or tested in other situations. The experiment reported here is such an attempt.

¹ This study was performed in part under Contract Nonr-248(55) and in part under Contract Nonr-4010(03) between the Office of Naval Research and The Johns Hopkins University. This is Report No. 1 under the latter contract. Reproduction in whole or in part is permitted for any purpose of the United States Government.

METHOD

The experimental task. The Ss in this experiment were told that they were preparing a tape for a digital computer and every attempt was made to lend credibility to this fiction. Specifically, the task required Ss to punch long sequences of random digits into a teletype tape by means of a teletype perforator. The digits to be copied were pointed on a roll of paper tape in two columns of five digits each. The spacings between the individual digits, the two columns, and the rows of digits were designed to make the numbers easily readable. They were presented in a simple exposure device and, at any one time, a total of 30 digits, three rows of 10 each, were visible in the opening of the exposure device. The S used his left hand to advance the paper tape on which the digits were printed. Both the paper tape on which the digits were printed, and the teletype tape into which the digits were punched, fed into reservoirs to which S did not have access. This made it impossible for S to estimate his output during a work session from the lengths of either tape.

The work was done in a room which housed a genuine, special-purpose digital computer, and the computer did, in fact, use tape of precisely the same kind as Ss were punching. This was readily apparent. A counter, mounted on top of the teletype perforator, could be activated, or deactivated, by a remotely-located switch. When it was activated, the counter tallied each stroke made on the perforator. Another counter, located in another room, was always operative, and all output data required for this experiment were taken from this second counter. A large wall-mounted clock, easily visible from the work table, completed the experimental arrangements.

Subjects. The Ss were 16 male undergraduate Hopkins students recruited through the student employment bureau and paid for their services. The only limitations placed on their selection were that they should (a) have good enough visual acuity to read the printed digits easily, (b) be right-handed (to operate the exposure device and the perforator in their fixed positions), and (c) be able to report for work an hour a day, every day of the week except Saturday and Sunday, at the same time of day. The Ss were told that they were being hired for an indefinite period of time (ostensibly "until the job was finished") but in actual fact were employed for a total of 24 days.

Instructions. To lend weight to the impression that this was a genuine work situation, instructions were neither memorized nor precisely standardized. However, the first set of instructions was always given by E and they included the following two points: (a) The Ss were told that they were preparing tapes for the computer in the room. The somewhat more sophisticated or inquisitive Ss who wanted to know why so many random digits were required were informed that they would be used for a Monte Carlo simulation. This was highly plausible considering the general nature of the research being carried out in the university at that time, and the explanation seemed to satisfy those Ss who inquired.

(b) Accuracy was not imperative, since these were random sequences and an occasional error would not affect the usefulness of the tape. The Ss were also told, however, that an excessive number of errors could not be tolerated. They were told otherwise to do as much as they could each day.

Each S was also instructed in the operation of each of the two principal items of equipment (the exposure device and the perforator) he would use, was given supplementary information about hours of work, pay, and so on, and then was left alone. In general, E never saw Ss again. They reported each day to a female research assistant and it was she who admitted them to the computer room, paid them weekly, and generally dealt with them after the first day. The research assistant was carefully instructed to play the role of a poorly informed assistant who did precisely what she had been instructed to do by a noncommunicative professor.

Experimental conditions. Each of the 16 Ss was assigned to one of four experimental groups, four Ss to a group.

Group I. These Ss never received any information about their output. The counter was never operative during their work sessions.

Group II. For Ss in this group the counter was operative, but the counter was never reset to zero, that is, there was always some reading in the counter at the start of the work session. Presumably, therefore, S in this group could, if he were so inclined, note the entry at the beginning of a work session, the reading at the end of a work session, and, by subtraction, determine his output for the period.

Group III. For Ss in this group, the counter was always operative and it was always set to zero before they reported for work.

Group IV. For Ss in this group, the counter was always operative and always set to zero before they reported for work. In addition, Ss in this group were asked to jot down on a card the reading on the counter at the end of 15 minutes, 30 minutes, and 45 minutes and to record their names and the date. They were informed that this record was merely for accounting purposes since it was a simple way for us to recall later that they had been present for work on the date alleged. At the end of an individual work session the card was collected by the research assistant and tossed into a drawer casually without any comment or other notice.

For Ss in the first three groups, neither E nor the research assistant called attention to the counter. It was merely there, in plain sight, for Ss to see or not, as they chose. Spontaneous comments directed toward the research assistant indicated that by the end of the fifth day of work, at least, every S in Groups II and III had noticed the counter and deduced that it was tallying strokes made on the perforator. For all four groups of Ss, the treatment was in other respects the same. Once S was admitted to the computer room and seated at his desk, he was never interrupted until the end of his session. No work quotas were set for him, nor was notice ever taken of the amount of work that he did.

Whenever *S* inquired about how much work he should do, the reply was always a casual "Do as much as you can." At all times, the research assistant attempted to create the impression that this was a genuine work task, that we were interested only in getting the work done, and that we were not at all interested in individual performances.

RESULTS

The raw data for this experiment are the numbers of digits punched at the end of 15 minutes, 30 minutes, and 45 minutes, taken from the remotely located counter. Although the work session was nominally an hour in length, in actual practice the sessions were a little less than that. Sessions were scheduled to begin on the hour but there was, of course, always some transition time involved in re-setting the counter (if this was required), meeting *Ss*, and so on. The three items of raw data are available for each *S* for each of 24 days.

The entire set of data was analyzed in a very large analysis of variance in which all the primary sources of variation (between the means of the 16 *Ss*, between the means for the four groups, between the means of *Ss* within groups, between the means for the three 15-minute segments of the work session, and between the means for the 24 days) and all of their associated interaction terms were tested. In addition, the variation between the means for the 24 days, and all the interactions involving days (seven in all), were subjected to a regression analysis using an extension of the method of orthogonal polynomials described by Grant (1956) and Anderson and Houseman (1942). In each of these cases, the linear, quadratic, cubic, quartic, and quintic components of the regressions were computed and tested for significance.

To present the table showing the full analysis of variance would require at least two pages and would add little to this article because the essential findings can be stated very simply: the variation due to the primary variable of interest in this study (the variation between groups) was not significant, nor were any of the interactions or regressions involving groups significant. This statement can be further amplified by saying that it was not a matter of "almost reaching" significance. The *F* ratios were generally so small that

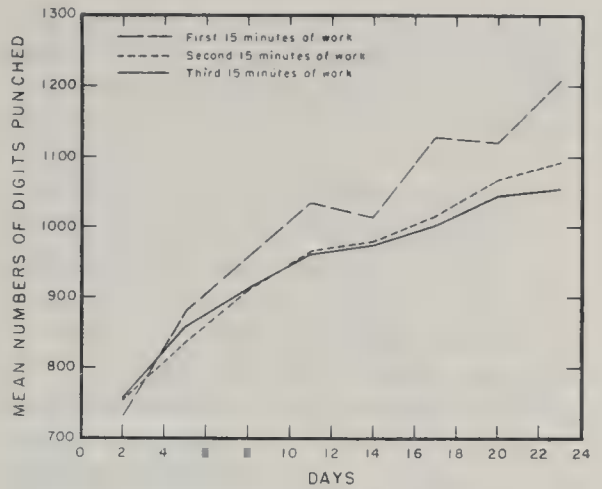


FIG. 1. Mean numbers of digits punched by 16 subjects during each 15-minute segment of the daily work session.

there was not even any reason to suppose that there were any trends worth examining further.

A few sources of variation were statistically significant, but these were incidental to the main purpose of the study. The most important of these was the variation between days for all the data combined. The regression analysis of these same data shows that the regression due to days has both a significant linear and quadratic component. The cubic and higher-order components of the regression are not significant. The general nature of the trend can be seen from Figure 1. Adding the three curves of Figure 1 together would give the overall trend of the data and reveal that performance increased steadily throughout the 24-day period, but that the regression was also significantly curvilinear.

The only other significant sources of variation in the entire analysis of variance are also revealed in Figure 1. They are: (a) significant differences among the means for the three 15-minute segments of each daily work session, (b) a significant interaction of 15-minute segments times days, and (c) significant differences among the linear components of the regressions of 15-minute segments as a function of days. On the average, *Ss* did more work in the first 15 minutes than they did in the two later 15-minute segments. This difference, however, was not apparent during the first few days and emerged as a consistent pattern only after about the fourth or fifth

day of work. This latter fact accounts for the significant interaction and for the significant differences among the regressions of the curves in Figure 1.

DISCUSSION

In terms of the primary purpose of this study, the results are disappointing. The trouble with negative findings is, of course, that they are open to at least two interpretations: (a) There is a real effect but the experiment designed to test it was not sensitive enough to reveal the effect, or (b) the null hypothesis is correct, that is, the variable under investigation has no real effect. Unfortunately, there is no easy way to decide between these two alternatives.

As regards the first alternative, however, it is important to note that I tested more Ss (16 versus 12), for more days per condition (24 versus 4), and for more total time per S per condition (24 hours versus 16 hours) than did Gibbs and Brown (1955) in their main experiment. Moreover, the most critical *F* ratio in this experiment had more degrees of freedom (3 and 12) available than was true in the Gibbs and Brown study (1 and 10). On the other hand, Gibbs and Brown had each S serve as his own control. This is generally a more sensitive type of experimental design than assigning Ss to independent groups as was done here. Adding further to the difficulties of interpretation is the fact that Gibbs and Brown appear to have made some basic errors in analyzing their data into component sources of variation. Taking all of these factors into account, it seems that in terms of purely statistical considerations this experiment was probably of the same order of sensitivity as the one performed by Gibbs and Brown. If there really is an effect here, and if the effect is as great as Gibbs and Brown report, then it seems reasonable to expect that this experiment should have discovered it.

The other plausible interpretation is that for the kind of experiment I have reported here, information per se does *not* serve as an incentive. What are some of the important differences between the Gibbs and Brown study and mine? There are, of course, many, but three, I believe, are particularly relevant to this discussion.

First is the difference between the total context of the two situations. Entirely aside from the particular task required of Ss (operating a copying machine versus a teletype perforator) is the total context which surrounded the task. Several statements in the Gibbs and Brown report allow one to state with some certainty that Ss in their experiment must have clearly perceived that it was only a laboratory experiment. The fact that half of Ss worked with the counter visible and the other half with the counter concealed and that *these two groups were then reversed*, suggests that Ss must have clearly perceived the context as an artificial work situation. In the study reported here every effort was made to create the impression that this was a real work situation. There is, however, no way of knowing how effective this stratagem was or whether Ss were indeed deceived. At any rate, it is at least likely that the total context of the two studies may have been sufficiently different to account for the difference in outcomes.

A second major difference between the two studies is that in the Gibbs and Brown experiment, two Ss worked on identical machines in the same room. In this study, each S worked individually. There is, therefore, the possibility that group competition, in addition to self-competition, contributed to the former. In addition, we cannot rule out certain other group effects in the Gibbs and Brown study—for example, the working tempos of the two men in each group may have been correlated, they may have taken rest pauses together, and so on.

The third major difference between the two studies is that Gibbs and Brown required their Ss (in the "knowledge of results" condition) to record the counter reading in a book "to ensure that the available information was being used." The men who had no knowledge of results made a tick in a book at the same intervals. In addition, the counter reading was photographed automatically at 10-minute intervals and this photographic record was made in the room where Ss were working. In my experiment, all recording of data was done from a remotely-located counter. There is, therefore, the strong possibility that whatever *Es* themselves state

about the situation, Ss perceived and thought that their performance was being recorded, studied, and compared in some way. To sum up, it may be that Gibbs and Brown did not really succeed in isolating the incentive aspect of knowledge of performance.

Two other differences in outcomes also serve to emphasize the differences between the two experiments. First, Gibbs and Brown report no learning effect in their experiment, that is, no change in daily performance over time. As has already been shown, this was the most significant source of variation in the study reported here. Second, Gibbs and Brown found no significant changes in performance throughout their 2-hour work periods. In this experiment there were such consistent changes in output.

To sum up, there appear to be some substantial differences between the Gibbs and Brown study and my own, and it is likely that the negative findings reported here might very well be traced to one or more of these sources. Or, conversely, the fact that Gibbs

and Brown obtained a positive effect may very well have been due to some additional factors present in their experiment which were not present in mine. At any rate, it seems clear that the precise circumstances under which knowledge of performance can serve as a pure incentive, if indeed it ever can, needs to be more clearly delineated.

REFERENCES

- AMMONS, R. B. Knowledge of performance: Survey of literature, some possible applications, and suggested experimentation. *USAF WADC tech. Rep.*, 1954, No. 54-14.
- ANDERSON, R. L., & HOUSEMAN, E. E. Tables of orthogonal polynomial values extended to $N = 104$. *Ia. State Coll. Agric. Exp. Sta. res. Bull.*, April 1942, No. 297.
- GIBBS, C. B., & BROWN, I. D. Increased production from the information incentive in a repetitive task. *Med. Res. Council, Appl. Psychol. Res. Unit, Great Britain*, 1955, (Mar.), No. 230.
- GRANT, D. A. Analysis-of-variance tests in the analysis and comparison of curves. *Psychol. Bull.*, 1956, 53, 141-154.

(Received August 9, 1963)

AN EXPERIMENTAL STUDY OF THE RELATION BETWEEN NURSING CARE AND PATIENT WELFARE¹

J. RICHARD SIMON AND WELLBORN R. HUDSON

State University of Iowa

4 experiments were conducted in an ongoing hospital setting to test the hypothesis that increasing the amount and quality of nursing care would produce improvements in patient welfare. Changes introduced on experimental wards consisted of increasing the size of the nursing staff, conducting inservice-education programs, and combining increased staffing and inservice education. With the exception of a reduction in patient complaints during 1 experiment, no improvement in patient welfare was produced. Results suggest that there is a limit to the contribution which nursing can make to patient welfare and that this limit is much closer to the existing level of care than was formerly thought.

This paper describes the results of a 3-year experimental study of the relationship between nursing care and patient welfare. Most nurses believe that a patient's welfare is directly related to the amount and quality of nursing care he receives. It was this belief which led to the hypothesis that *increasing* the amount and quality of nursing care would produce improvements in patient welfare.

Four experiments were conducted during which changes in nursing care were introduced on experimental wards in the State University of Iowa Hospitals. These changes included increasing the nursing care hours available per patient and instituting various inservice-education programs for nursing service personnel. The criterion for evaluating these changes was the welfare of the patients on the wards. It is this use of patient welfare as the criterion for evaluating changes in nursing which differentiates this study from most other research in nursing.

METHOD

Criterion Measures

Researchers have long recognized that the welfare of the patient is the most relevant criterion for

evaluating changes in nursing practice. However, patient welfare is difficult to define and even more difficult to measure. Therefore, one of the major problems in this research was to develop reliable and sensitive instruments for measuring patient welfare.

The first step toward solving the criterion problem was to ask a large number of medical and nursing authorities to identify those physical and behavioral characteristics which reflected a patient's welfare. The next step was to construct rating scales and other formal means of measuring these various dimensions of patient welfare. Only a brief description of the measures is provided here since a detailed account of their development is available elsewhere (Nurse Utilization Project Staff, 1960).

One group of patient welfare measures used in this research included objective indices such as length of hospital stay, number of postoperative days, number of fever days, and doses of narcotics, analgesics, and sedatives. If the level of patient welfare increased on a ward, it seemed reasonable that this change would be reflected by a decrease in the average number of hospital days, postoperative days, and fever days, as well as by a change in the number, dosage, or kind of drugs given.

A second group of measures included nurses' ratings of the patients' mobility status, mental attitude, physical independence, and skin condition (Simon, 1961a; 1961b). Separate rating scales were constructed to measure each of these aspects of patient welfare. The mobility measure described the degree to which a patient had resumed the normal physical activities of sitting, standing, and walking. The mental attitude measure described the patient's observable verbal and emotional responses to his environment and to his treatment. The physical independence measure described the degree to which the patient was dependent upon the nurses for his physical care in three areas; bathing, nutrition, and elimination. The skin condition measure described the number, size, and seriousness of skin lesions. Patients were rated daily on mobility, mental attitude, and physical independence whereas skin condi-

¹ This research was supported by Research Grant GN-4786; National Institutes of Health; United States Public Health Service, and by the State University of Iowa. Principal Investigators were Myrtle E. Kitchell Aydelotte and Marie E. Tener. The authors provided technical direction of the project. Nurse members of the research staff were: Sally Chastain, Jeanette Hoffman, Pearl Johnson, Jane Kroetsch, Marian Olson, June Rekwart, Elizabeth Sprague, Frances Walker, and Dolores Whitehead.

TABLE 1
AVERAGE DAILY NURSING CARE HOURS, PATIENT CENSUS, AND NURSING CARE HOURS
PER PATIENT: INCREMENT EXPERIMENT: MEDICINE

		Nursing care hours		Patient census	NCH/P
Data period		M	SD		
Control ward	Base 1	66.38	3.58	27.6	2.42
	Base 2	63.97	3.71	21.5	3.00
	Base 3	64.33	1.67	23.4	2.78
Experimental ward	Base	66.94	4.35	24.9	2.69
	Increment 1	90.99	4.01	18.2	5.05
	Increment 2	91.92	2.41	24.6	3.78

tion ratings were made every other day. The intraclass correlation coefficient (Ebel, 1951) was used to obtain conservative estimates of the reliability of the ratings. These ranged between .73 and .87 (Simon, 1961a).

Another group of ratings was provided by the senior ward resident. He evaluated each patient's general condition on admission to the hospital and again on discharge so that any change in general condition could be computed. The physician also rated the patient's response to treatment throughout his hospital stay.

A structured patient-opinion questionnaire was used to assess the psychogenic aspect of patient welfare. Each patient was interviewed twice concerning the treatment or care he was receiving. The number of complaints made during these structured interviews was taken as an index of the amount of worry, concern, or anxiety that the patient felt.

Still other indices of patient welfare were provided by a behavior-sampling procedure (Simon & Chastain, 1960; Simon, 1960). Each patient was observed 24 times a day during each day of his hospital stay, and his activity was recorded in various predefined categories. The rationale behind this approach was that a patient's welfare is reflected in the way he spends his time in the hospital. Four indices were developed from these sampling data: time (%) spent in bed, time (%) spent in chair, time (%) spent up, and time (%) spent in communication and occupied leisure activities.

Experimental Design

Four experiments were conducted in which the nursing staff on an experimental ward was changed in some way. The changes consisted of either increasing the size of the nursing staff, conducting an inservice education program for nursing service personnel, or combining increases in ward staffing and an inservice education program. The welfare of the patients on the experimental ward was then compared with comparable data from a control ward or with data from the same ward prior to the time the experimental changes were introduced.

Data periods varied in duration from experiment to experiment and were between 4 and 6 weeks long. Patient welfare data were collected on all patients who were admitted to the ward after a data period started, were discharged before the data period ended, and were in the hospital at least 4 days. The number of patients involved in the four experiments, listed in chronological order, was 111, 124, 154, and 96.

RESULTS

Increment Experiment: Medicine

This experiment was conducted on two female medical wards for three 4-week periods. Since the wards were comparable in terms of physical environment, patient population, and composition of the nursing staff, one ward served as the experimental ward and the other as the control. During the first 4 weeks, patient welfare data were collected under base or normal staff conditions on both wards. During the second and third 4-week periods, the number of nursing hours available on the experimental ward was increased by approximately 35% over the base period while the staff on the control ward was maintained at so-called normal level. The increase in staff on the experimental ward was achieved mainly by adding graduate nurses on both the 7-3 and 3-11 shifts. No increment was made on the 11-7 shift.

Table 1 shows the average daily nursing care hours during each data period.² The

² In computing the total number of nursing care hours, a sophomore student hour was arbitrarily counted as equivalent to one-fourth of a junior, senior, or graduate nursing hour. The hours of subsidiary personnel were counted the same as graduate hours.

small standard deviations indicate the close control maintained over the nursing time available each day. Comparable control over the daily patient census was not possible,³ and consequently the data periods differed in terms of average nursing care hours available per patient (NCH/P). (See Table 1). For example, during Increment 1, the NCH/P was 88% higher than during the base period while the comparable figure for Increment 2 was 41%. This difference between the increment periods was due to a drop in the average census during Increment 1.

In addition to collecting patient welfare data throughout this experiment, systematic observations were also made of the activity of the ward nursing staff during each data period. The daily sampling of the way that ward personnel spent their time indicated that, during the increment periods, the average patient received up to 60% more direct nursing care than she received during the base periods. Despite this sizable increase in direct patient care, *t* tests indicated that, with one major exception, there was no significant improvement in patient welfare during the increment periods. The only indication that incrementing the nursing staff had a favorable effect was a significant reduction (*p* < .05) in the number of complaints patients made during the structured interviews. An analysis of individual questionnaire responses indicated that during the periods of incremented staffing, patients were happier with the quality, amount, and promptness of care which they received.

Combination Increment and Training Experiment: Surgery

In this experiment, two variables were manipulated simultaneously in an attempt to improve patient welfare. One variable was incrementing; i.e., additional graduate nurses were again added to the ward staff in order to increase the amount of care each patient received. The other variable, inservice education, was introduced in an attempt to improve the quality of nursing care given.

³ The two wards used were the only female medical wards in the hospital, and the admission policy of the hospital could not be altered.

The experiment was conducted on a male surgical ward over a 21-week period. (No comparable surgical ward was available as a control.) During the first 6 weeks, patient welfare data were collected at base level of staffing. This base period was followed by a 3-week inservice education program during which an incremented nursing staff attended formal lectures and took part in informal clinical instruction. Following the inservice-education program was a 6-week period during which the level of staffing was approximately 40% above base level. Then followed another 6-week period during which the staff was reduced to approximately the original base level.

The inservice education program was planned to fit the particular needs of the staff and included: a review of basic nursing care and surgical nursing practice; training in observation and evaluation; discussions of interpersonal relations with particular emphasis on nurse-patient relations; and introduction of a team nursing plan of patient care (which was continued for the remainder of the experiment). All instruction sessions were tape recorded and repeated for those staff members who were unable to attend the original sessions.

Table 2 shows the average daily nursing hours, patient census, and NCH/P during the three data periods. The NCH/P during the base, posttraining period was 8% higher than during the base, pretraining period because of a slight drop in the patient census. Analysis of the pattern of nursing activity indicated that patients received 28% more direct patient care during the increment period than during the base. When the staff size

TABLE 2
AVERAGE DAILY NURSING CARE HOURS, PATIENT CENSUS, AND NURSING CARE HOURS PER PATIENT: COMBINATION INCREMENT AND TRAINING EXPERIMENT: SURGERY

Data period	Nursing care hours		Patient census NCH/P	
	<i>M</i>	<i>SD</i>		
Base, pretraining	84.71	5.19	28.0	3.04
Increment, posttraining	115.40	6.88	27.7	4.18
Base, posttraining	83.81	2.35	25.8	3.27

TABLE 3
AVERAGE DAILY NURSING CARE HOURS, PATIENT CENSUS, AND NURSING CARE
HOURS PER PATIENT: TRAINING EXPERIMENT: MEDICINE

		Nursing care hours		Patient census	NCH/P
		<i>M</i>	<i>SD</i>		
Control ward	Base 1	67.69	3.34	22.5	3.02
	Base 2	68.45	4.02	22.4	2.85
Experimental ward	Base	70.71	5.90	25.7	2.82
	Posttraining	68.67	5.87	21.0	3.40

returned to normal during the final 6-week period, the amount of direct patient care time also dropped to the normal level.

Analysis of the patient welfare scores collected during the experiment failed to disclose any significant differences between data periods. Still, there was an almost unanimous feeling among both nurses and physicians that things were somehow "better" following the inservice-education program. None of the staff, however, could offer any concrete evidence to support this feeling. Nursing experts suggested that the 3-week training program may have been too short to produce any marked change in the level of patient welfare and recommended that subsequent experiments investigate the effect of a more extensive inservice education program. Before the final two experiments were conducted, the existing patient welfare measures were revised in an effort to increase their sensitivity, and additional measures were also introduced.⁴

Training Experiment: Medicine

This experiment was conducted on the two wards previously used for the increment experiment: medicine. During the first 6-week period, both the experimental and control wards were observed at the base level of staffing. During the second 6-week period, the staff on the experimental ward participated in an inservice education program. Following this program was another 6-week data period. Table 3 shows that the data periods were similar in nursing care hours available

but, because of a fluctuating census, differed somewhat in NCH/P.

The inservice education program was planned, coordinated, and conducted by a full-time nurse preceptor who spent 2 weeks on the ward determining the specific needs of the staff before the training period began. The content of the inservice education program is described in detail elsewhere (Nurse Utilization Project Staff, 1960). It consisted of lectures, conferences, demonstrations, and individual instruction covering the general areas of medical theory and therapy, nursing care, ward management, and interpersonal relations.

In order to determine whether mean patient welfare scores differed from one data period to another, a Treatment \times Levels analysis of variance technique was used. It had been observed that most of the measures of patient welfare were correlated with both the patient's age and his general condition on admission rating. Therefore, by assigning patients to one of four different levels on the basis of their age and general condition, the variance attributable to differences between levels could be identified and eliminated from the error term, thus permitting a more precise test for differences between data periods. There were 32 patients in each of the four data groups, 26 subjects being discarded in order to fulfill the proportionality requirement of the Treatment \times Levels design.

Table 4 summarizes the Treatment \times Levels analyses of nine measures of patient welfare. None of the analyses provided any evidence that inservice education produced an improvement in patient welfare. There was a sta-

⁴ The indices based on random sampling of the patients' behavior were not used in the first two experiments.

TABLE 4
SUMMARY OF ANALYSES OF VARIANCE OF PATIENT WELFARE MEASURES:
TRAINING EXPERIMENTS: MEDICINE AND UROLOGY

Patient welfare measures	Training experiment: medicine			Training experiment: urology		
	Levels	Treatment × Levels	Treat- ment	Levels	Treatment × Levels	Treat- ment
Hospital days	*	<i>ns</i>	*	<i>ns</i>	<i>ns</i>	<i>ns</i>
Mobility	**	<i>ns</i>	**	**	<i>ns</i>	<i>ns</i>
Mental attitude	**	**	**	**	<i>ns</i>	<i>ns</i>
Physical independence	**	<i>ns</i>	<i>ns</i>	**	<i>ns</i>	<i>ns</i>
Skin condition	*	<i>ns</i>	<i>ns</i>	**	*	<i>ns</i>
Time (%) in bed	**	<i>ns</i>	<i>ns</i>	**	<i>ns</i>	<i>ns</i>
Time (%) in chair	**	<i>ns</i>	<i>ns</i>	**	<i>ns</i>	<i>ns</i>
Time (%) up	<i>ns</i>	<i>ns</i>	<i>ns</i>	**	<i>ns</i>	<i>ns</i>
Time (%) communication and occupied leisure	**	<i>ns</i>	*	**	<i>ns</i>	<i>ns</i>

* $p \leq .05$.
** $p \leq .01$.

tistically significant difference between data periods on four measures; hospital days, mobility, mental attitude, and time (%) in communication and occupied leisure, but, in none of these cases, did the posttraining data period have the highest mean score. The reasons for these differences are not clear, especially since the four data periods were comparable with respect to age and general condition of the patients.⁵ In eight of the nine analyses, the levels effect was significant indicating that the assignment of patients to levels eliminated a significant source of variation. In only one instance, for the mental attitude measure, was there a significant Treatment × Levels interaction.

The patient complaints data for the three base periods were pooled and compared with similar data from the posttraining period. The average number of complaints during the base periods was 9.4 as opposed to 7.8 for the posttraining period. Although this difference was in the expected direction it was not statistically significant ($F = 2.68$; $p > .10$). None of the remaining measures (i.e., fever days, response to treatment, change in general condition, dosage of drugs,

⁵ An analysis of the first day's ratings indicated that these differences between data periods in overall mean ratings are not a result of chance differences in the level of patient welfare at the time of admission.

etc.) gave any indication that the inservice-education program improved patient welfare.

Training Experiment: Urology

The purpose of this final experiment was to determine the effect of a 6-week inservice-education program on the welfare of male urological patients. The training program, while specially designed to meet the needs of the ward staff, was quite similar to that of the previous experiment. Six-week data periods preceded and followed the inservice education program. Table 5 shows that the data periods were almost identical in terms of nursing hours available, average census, and NCH/P.

Treatment × Levels analyses of variance were again used to test for differences in patient welfare between data periods. There

TABLE 5
AVERAGE DAILY NURSING CARE HOURS, PATIENT CENSUS, AND NURSING CARE HOURS PER PATIENT:
TRAINING EXPERIMENT: UROLOGY

Data period	Nursing care hours		Patient census NCH/P	
	<i>M</i>	<i>SD</i>		
Base	65.70	6.42	22.2	3.01
Posttraining	68.43	6.95	22.6	3.08

were 47 patients in each treatment group (two having been discarded from the pre-training group to maintain proportionality). Table 4 summarizes the analyses of variance. In none of the analyses was there a difference between the pre- and posttraining periods. The levels factor was significant in eight of the analyses, and in one analysis, skin condition, there was a significant Treatment \times Levels interaction. Analysis of the patient complaints data indicated a tendency toward fewer complaints during the post-training period but the difference was not significant. The remaining patient welfare measures also failed to show any differences.

DISCUSSION

Results of these experiments may be summarized as follows: no improvement in patient welfare was produced by substantially increasing the size of the ward staff, by conducting inservice-education programs, or by combining staff increases and inservice education. The only exception was a significant decrease in patient complaints during the period of incremented staffing on the Medical Service.

There are several possible explanations for these results. The most reasonable explanation is that the nursing staff was already making its maximum contribution to patient welfare before any changes were introduced. To use an analogy, regular watering of a lawn will make it greener. However, beyond a certain point, additional water is, for all practical purposes, wasted. Results of this study suggest that there is a limit to the contribution which nursing can make to patient welfare and that this limit is much closer to the existing level of care than was formerly thought.

As an alternative explanation, one might argue that patient welfare did, in fact, improve but that the criterion measures were not sensitive enough to detect a change. This, however, seems highly unlikely since the measures were sensitive enough to: enable raters to make reliable discriminations between individual patients; detect significant differences between groups of patients assigned to different levels based on their age and general condition; detect improvement in

some groups of patients during their stay in the hospital; and detect differences between experimental data periods (though not in the hypothesized direction). There is considerable evidence, therefore, that if a change had occurred, it would have been detected (Simon, 1961a).

Another possibility is that the favorable effect of the experimental variables may have been concealed by the action of some uncontrolled variable(s). This explanation is also unlikely since all obviously relevant variables were either controlled or their possible effects investigated. For example, whenever a control ward was not available, additional analyses were undertaken to ensure comparability of the patient populations during the different data periods.

A third alternative explanation for the failure to improve patient welfare is that the experimental variables did not sufficiently change either the amount or the quality of patient care. From a purely practical viewpoint, however, it is doubtful whether most hospitals could afford either to increase patient care beyond the level achieved in this study or to provide more comprehensive inservice education programs.

Why, if there was no change in patient welfare, did many nurses feel that things were somehow "better" during the experimental periods? At least two mechanisms might be involved. First, the experimental changes resulted in the nurses being better informed about their patients, which, in turn, may have led them to feel that their patients were better off. Secondly, if a nurse gave better patient care or at least felt that she was providing better care, it would be quite reasonable for her to believe that her patients were better off—even though she could not identify any concrete evidence to support her belief. We prefer to believe that the reason things seemed better was not that the patients had changed, but instead, that the nurses had changed.

What are the implications of our findings? First, existing inservice education programs should be reevaluated to determine whether they are actually doing the job they are assumed to do. Second, hospitals should not assume that adding more nurses will neces-

sarily improve patient welfare. Before making any definitive statement concerning the optimum level of staffing, we would want to see results of similar studies performed in other hospitals where the normal level of staffing is considerably lower than in the present study. Perhaps under these conditions, increasing the staff would produce improved patient welfare. It is also possible that a younger patient population would have been more responsive to the experimental variables. One point is clear, however; the relationship between nursing care and patient welfare is not as direct as most nurses have heretofore believed.

REFERENCES

- EBEL, R. L. Estimation of the reliability of ratings. *Psychometrika*, 1951, **16**, 407-424.
- NURSE UTILIZATION PROJECT STAFF. *An investigation of the relation between nursing activity and patient welfare*. Iowa City: State Univer. Iowa, 1960.
- SIMON, J. R. Patient activity as a measure of patient welfare. *Hosp. Mgt.*, 1960, **90** (3), 95-100.
- SIMON, J. R. Nurses' ratings of patient welfare as criterion measures in the health sciences. *Occup. Psychol.*, 1961, **35**, 10-22. (a)
- SIMON, J. R. Systematic ratings of patient welfare. *Nurs. Outlook*, 1961, **9**, 432-436. (b)
- SIMON, J. R., & CHASTAIN, SALLY S. Take a systematic look at your patients. *Nurs. Outlook*, 1960, **8**, 509-512.

(Received August 12, 1963)

*A pioneering study
in industrial psychology*

MANPOWER DEVELOPMENT

by **ELIAS H. PORTER**

*Senior Human Factors Scientist,
System Development Corporation*

The System Training Concept, originated for large-scale Air Force operations, was a remarkable demonstration of how crews can achieve high efficiency under stress in complex situations. Here one of the scientists behind its development describes this concept and extends it for application to organizations of all kinds: schools, businesses, national defense, and municipal protection agencies.

"A significant contribution to our understanding of the problems of manpower development."—JOHN L. KENNEDY, *Chairman, Department of Psychology, Princeton University*

\$4.95

. . . and a unique handbook

EMOTIONAL HEALTH: In The World of Work

by **HARRY LEVINSON**

*Director, Division of
Industrial Mental Health,
The Menninger Foundation*

Here is the first book to apply sound principles of psychology to emotional disturbance in the work environment. Addressing the executive, Dr. Levinson discusses problems that may result from the pressures of professional rivalry, personality clashes, age and retirement, etc., and the symptoms by which these problems can be recognized. The book is illustrated with actual case studies. "Particularly timely . . . a magnificent job."—ALFRED J. MARROW, *Chairman of the Board, Harwood Manufacturing Corporation*

\$6.95

At your bookstore or from Dept. 32



Harper & Row 49 E. 33rd St.
New York 10016

MANAGEMENT SCIENCE

QUARTERLY JOURNAL OF THE INSTITUTE OF
MANAGEMENT SCIENCES



*Make TIMS your information center for
new ideas in the Management Sciences*



For membership and subscription information write to

THE INSTITUTE OF MANAGEMENT SCIENCES

Box 626

Ann Arbor, Michigan

JOURNAL OF INDUSTRIAL PSYCHOLOGY

Edited by LEE W. COZAN

Quantitative investigations in industrial psychology and personnel management: job and worker analysis, employee performance evaluation, employment interviewing, psychological testing, work conditions and productivity, employee and supervisory training, and worker monotony, motivation and morale.

Published quarterly beginning February 1963

Annual Subscription: \$8.00 (Overseas \$9.00)

\$2.50 per issue

Send Orders and Manuscripts to

ELIAS PUBLICATIONS

P. O. Box 662, Washington 4, D. C.

W. H. RAYGROVE COLLEGE LIBRARY
DETROIT, MICHIGAN
PLEASE DO NOT REMOVE

Journal of Applied Psychology

KENNETH E. CLARK, Editor
University of Rochester

Table of Contents

Sensory-Feedback Analysis of Reading . . . Karl U. Smith, Richard Cambria, and James Steffan	275
The Attitudes of Research Chemists John R. Hinrichs	287
An Investigation of the Criterion Problem for a Medical School Faculty Calvin W. Taylor, Philip B. Price, James M. Richards, Jr., and Tony L. Jacobsen	294
Estimating the Number of Different Selection Decisions Resulting from the Use of Alternate Predictor Composites Warren W. Willingham	302
Job Attitudes in Management: VI. Perceptions of the Importance of Certain Personality Traits as a Function of Line versus Staff Type of Job Lyman W. Porter and Mildred M. Henry	305
A Study of Attitude Change in the Preretirement Period Shoukry D. Saleh	310
Convergent and Discriminant Validity for Areas and Methods of Rating Job Satisfaction Edwin A. Locke, Patricia Cain Smith, Lorne M. Kendall, Charles L. Hulin, and Anne M. Miller	313
Relationships between Job Difficulty, Employee's Attitude toward His Job, and Supervisory Ratings of the Employee Effectiveness . . Byron Svetlik, Erich Prien, and Gerald Barrett	320
An Analysis of Vocational Interests at Two Levels of Management Hrach Bedrosian	325
The Analysis of Job Performance by Multidimensional Scaling Techniques Douglas G. Schultz and Arthur I. Siegel	329
Use of a Logically Related Predictor in Determining Intragroup Differential Predictability John H. Steinemann	336

American Psychological Association

Consulting Editors

GEORGE E. BRIGGS, *Ohio State University*
MARVIN D. DUNNETTE, *University of Minnesota*
NORMAN FREDERIKSEN, *Educational Testing Service*
LEONARD D. GOODSTEIN, *University of Cincinnati*
EDWIN R. HENRY, *Peace Corps*
JOHN HOLLAND, *American College Testing Program*

CLIFFORD E. JURGENSEN, *Minneapolis Gas Company*
LAURENCE S. MCGAUGHRAN, *University of Houston*
QUINN MCNEMAR, *Stanford University*
HAROLD F. ROTHE, *Beloit Corporation*
THOMAS A. RYAN, *Cornell University*
ALEXANDER G. WESMAN, *Psychological Corporation*
CLARK L. WILSON, *Harvard Business School*

This journal gives primary consideration to original investigations in any field of applied psychology except clinical psychology, although a descriptive or theoretical article may be accepted if it represents a special contribution in an applied field. Quantitative investigations of interest or value to psychologists working in the following broad fields will be considered: vocational and educational prognosis, diagnosis, and guidance at the secondary and college level; personnel research in business, industry, and government; engineering psychology; industrial working conditions; research on opinion and morale factors; job analysis and classification research; market and advertising research.

Manuscripts must be accompanied by an abstract of 100–120 words typed on a separate sheet of paper. The abstract should conform to the style of *Psychological Abstracts*. Detailed instructions for preparation of the abstracts appeared in the *American Psychologist* (1961, 16, 833), or they may be obtained from the Editor or from the APA Central Office.

Manuscripts should be addressed to the Editor:

Dr. Kenneth E. Clark
College of Arts and Science
302 Morey Hall
University of Rochester
Rochester, N. Y. 14627

All manuscripts must be submitted in duplicate. Original figures are prepared for publication; duplicate figures may be photographs or pencil-drawn copies.

Manuscripts must conform to the style requirements described in the *Publication Manual of the American Psychological Association*.

The following policies govern reprinting of materials copyrighted in APA journals: (a) to require approval to reprint for tables and figures, and for text only if more than 500 words [total from one article] in length; (b) to continue approval to reprint, contingent on the author's approval, for articles reprinted in whole or in major part; (c) to negotiate where possible the dedication of royalties from commercial publishers to the American Psychological Foundation.

Journal of Applied Psychology

Published bimonthly by the
American Psychological Association
Prince and Lemon Sts., Lancaster, Pa. 17604
and 1333 Sixteenth St. N. W.
Washington, D. C. 20036

\$10.00 per volume

\$2.00 per issue

HELEN ORR
Managing Editor

ELIZABETH S. REED
Advertising Manager

FRANCES L. BREWER
Editorial Assistant

Subscriptions, orders, and business communications should be addressed to the American Psychological Association, 1333 Sixteenth St. N. W., Washington, D. C. 20036. Address changes must reach the subscription office by the tenth of the month to take effect the following month. Undelivered copies resulting from address changes will not be replaced; subscribers should notify the post office that they will guarantee second-class forwarding postage. Other claims for undelivered copies must be made within four months of publication.

Second class postage paid at Lancaster, Pa. and at additional mailing offices.

© 1964 by the American Psychological Association, Inc.

Journal of Applied Psychology

VOL. 48, No. 5

OCTOBER 1964

SENSORY-FEEDBACK ANALYSIS OF READING

KARL U. SMITH,¹ RICHARD CAMBRIA, AND JAMES STEFFAN

University of Wisconsin

This research investigates the space displacement of printed matter as an experimental technique for sensory-feedback analysis of reading behavior. 1 experiment compares the relative effects of inversion and reversal of reading material on errors and time of reading. Another determines angular rotation breakdown displacement angles of normal reading pace. Results show that reversal and inversion of the visual feedback of reading have more or less equivalent effects on the overall efficiency of reading performance. Significant asymmetric breakdown angular-displacement thresholds for right and left rotation of printed material also have been found. These differential-rotational thresholds may be related to handedness. The results suggest that the anisotropy of reading behavior is based on movement specialization related to the dynamic spatial orientation of eye movements and of the visual sensory data generated by such movements.

In this paper we describe certain new experimental approaches to the problem of reading which relate to sensory-feedback techniques of analyzing educational skills, and other patterns of human behavior (Smith & Smith, 1962). Prior psychological and educational studies of reading generally have been concerned with such static concepts as attention, recognition, and memory storage, or with the description of eye movements in reading. This latter type of research is one of the oldest of scientific motion analyses.

The well-known motion pattern of reading, a series of fixations alternating with saccadic jerks, gives us a basis of comparing reading efficiency in different individuals, and with different types of material, but it does not reveal much about the stimulus-response relationships that determine the organization of reading behavior and other aspects of form perception. A basic property of the perceived material is that it is anisotropic, i.e., it assumes

different configurational features when oriented in different spatial directions. The phenomenon of anisotropy, in perception, which was first delineated by Mach (1897) late in the nineteenth century, is a main aspect of response dimensionality that serves as the starting point in our studies of the spatial organization of motion. We do not consider behavioral anisotropy or dimensionality a limited phenomenon of form perception, but a primary feature of somatic behavior.

Part of the present experiment has been the technical task of devising new techniques to test the anisotropic characteristics of reading in terms of dynamic sensory-feedback relationships. Thus, we have explored the use of special television circuits, optical systems, and motor-driven display devices, in order to control the space displacement of different types of visual material. In the experiments to be reported, we have produced systematic disorientation of letters and words in various spatial axes and dimensions, i.e., inverted, reversed, both inverted and reversed, rotated, and angularly displaced visual feedback of the reading matter (Smith & Smith, 1962). The original efforts in research on this problem of

¹This experiment's funds came from a National Science Foundation grant for a project on Sensory-Feedback Analysis of Human Motion (NSF-3785) and a National Institutes of Health grant for a project on Brain Biochemistry, Development and Aging (MH-4785).

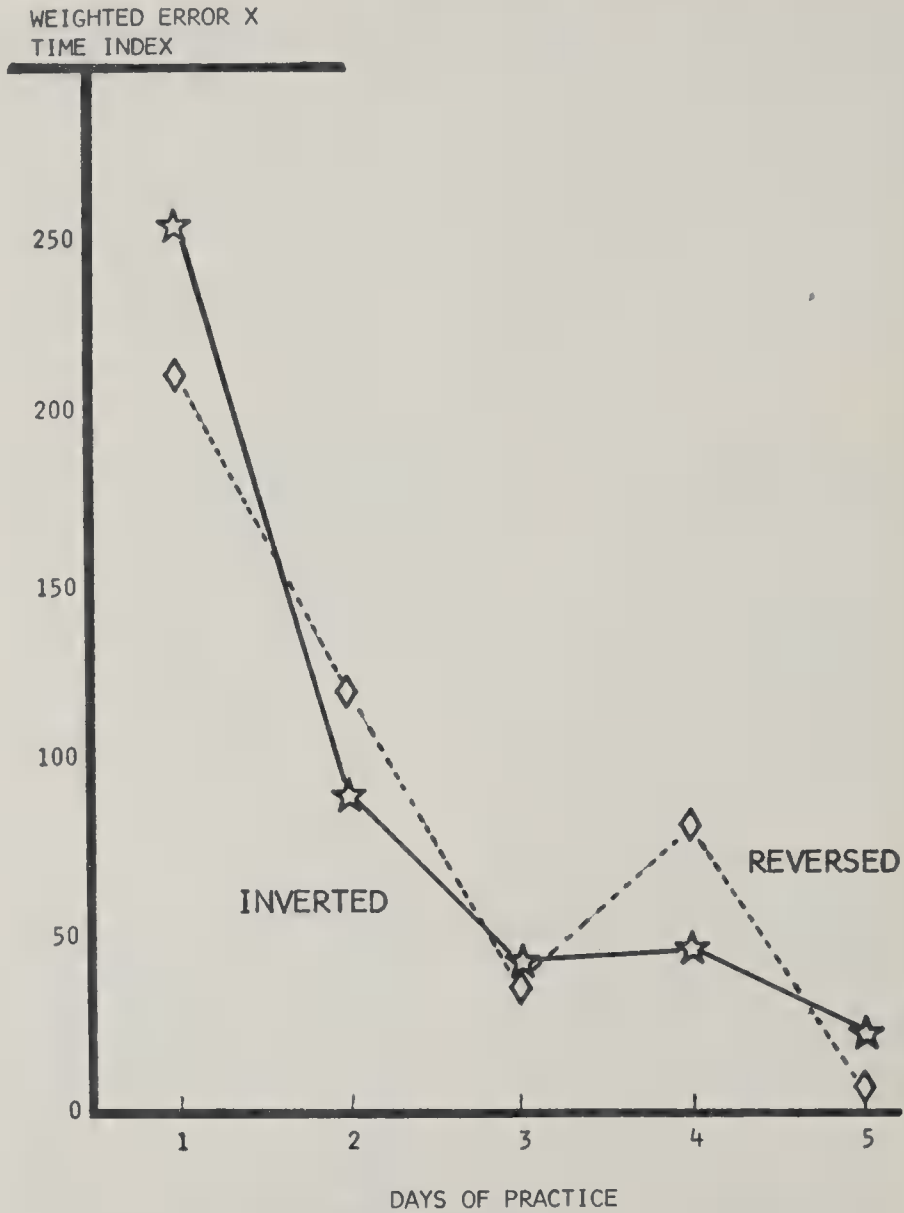


FIG. 1. Curves of learning for weighted Time \times Error scores for inverted and reversed orientations of the printed matter.

the effects of space displacement of visual feedback on reading were defined by Rhule and Smith (1959).

The theory guiding this research revises conventional theory that defines reading entirely as a learned verbal skill and the sequential linking of reinforced learned elemental responses as the basis of its integration (Mowrer, 1963). The principal assumption tested in this study is that the anisotropic organization of reading and of form perception in general is intrinsic to the sensory-feedback mechanisms of eye movements and their interactions with the general transport

and postural movement mechanisms of the body.

In the present research, four experiments dealing with space displacement of the visual feedback of reading have been carried out. The first determined the relative effects of inversion and reversal of printed matter on the time and errors of reading. The second and third determined the effects of the relative angular displacement of reading matter in a plane normal to vision. The fourth dealt with both systematic inversion and reversal and with angular displacement. The predictions for the outcome of these studies are that the

inversion and reversal of the reading will produce significantly distinctive patterns of performance, and that, with the angular rotations, a range of normal angular displacement and breakdown angles of such displacement will be found.

EXPERIMENT I: COMPARATIVE EFFECTS OF INVERSION AND REVERSAL OF READING²

In the first experiment the effects of inversion and reversal of the visual feedback of reading single letters, numbers, and combinations of characters were compared under controlled conditions. The material consisted of approximately 500 upper-case letters and numbers typed on bond paper. All letters of the alphabet and all numbers from one to nine were used. The letters and numbers were divided into five separate groups and ordered for difficulty on each page. The order consisted of single letters, single numbers, groups of two letters, groups of one letter and one number, and groups of three letters.

Two groups of four subjects (Ss) were used. One group learned inverted reading and the other learned reversed. Both groups were trained for a period of 5 days, one each of which a complete letter and number series was given; the particular series was different each day. On the first and last days, transfer tests were given. That is, the group learning inverting reading was given a reversed pretest on the first trial in addition to the inverted test, and on the last trial a reversed posttest was given in addition to the inverted test. A similar procedure was used for the reversed condition, except that the pre- and posttests were with inverted feedback.

The results of these preliminary experiments are summarized in terms of the difference in combined error/time scores. These scores were computed in terms of a weighted Standard Deviation Time \times Error score. As shown in Figure 1, the two learning functions for the inverted and reversed conditions are much the same. Thus, unlike the differences between inverted and reversed visual feedback that we typically see in manual motion, the effects of these two conditions of feedback on

eye movements in reading are very similar. This finding has been confirmed in later studies in which inversion, reversal, and angular rotation of printed matter have been compared.

The transfer effects in this experiment correspond in general to the learning findings. The percentage changes in time and error performance for the reversed and inverted transfer effects were computed separately. There was no difference between the percentage transfers for the inverted and reversed displacement conditions in terms of time scores. But the inverted transfer effect on errors was somewhat greater than the reversed transfer effect on errors in inverted reading.

EXPERIMENT II: ROTATIONAL BREAKDOWN THRESHOLDS

This experiment tested the assumption of neurogeometric theory of form perception that there are specific angular displacement breakdown thresholds for different directions and patterns of printed matter. The specific assumption is that there are differential ranges of angular displacement involved in the normal control of eye movements in reading. First, there is a normal range of displacement within which the eye-movement pattern will not be disturbed by the displacement. At the limit of this normal range, there is a breakdown threshold angle at which reading activity will be disorganized in a limited way; and



FIG. 2. Rotating reading pacer used to measure right and left rotational breakdown thresholds in reading.

² Jeffry Arbetmann assisted in this part of the experiment.

there is a breakdown range within which the degree of disturbance of motion will vary as a function of the magnitude of the displacement angle. It was anticipated that the breakdown thresholds for rotational displacement to the left and to the right would differ for the right-handed Ss.

Method

The special reading rotator shown in Figure 2 was used in the study. This consists of a motor-driven

reading pacer mounted on a rotating disk. The pacer could be set by S to correspond to his rate of reading out loud at a normal pace. The disk on which this pacer was located was turned by means of a larger motor system. Direction of rotation of the disk and pacer could be reversed by adjusting the pulley.

The task of S in this situation was to continue reading while the disk and reading pacer slowly rotated. The rate of rotation was set at roughly 180 degrees per minute. The S viewed the material without turning his head and read out loud as the disk and pacer turned. When the point was reached

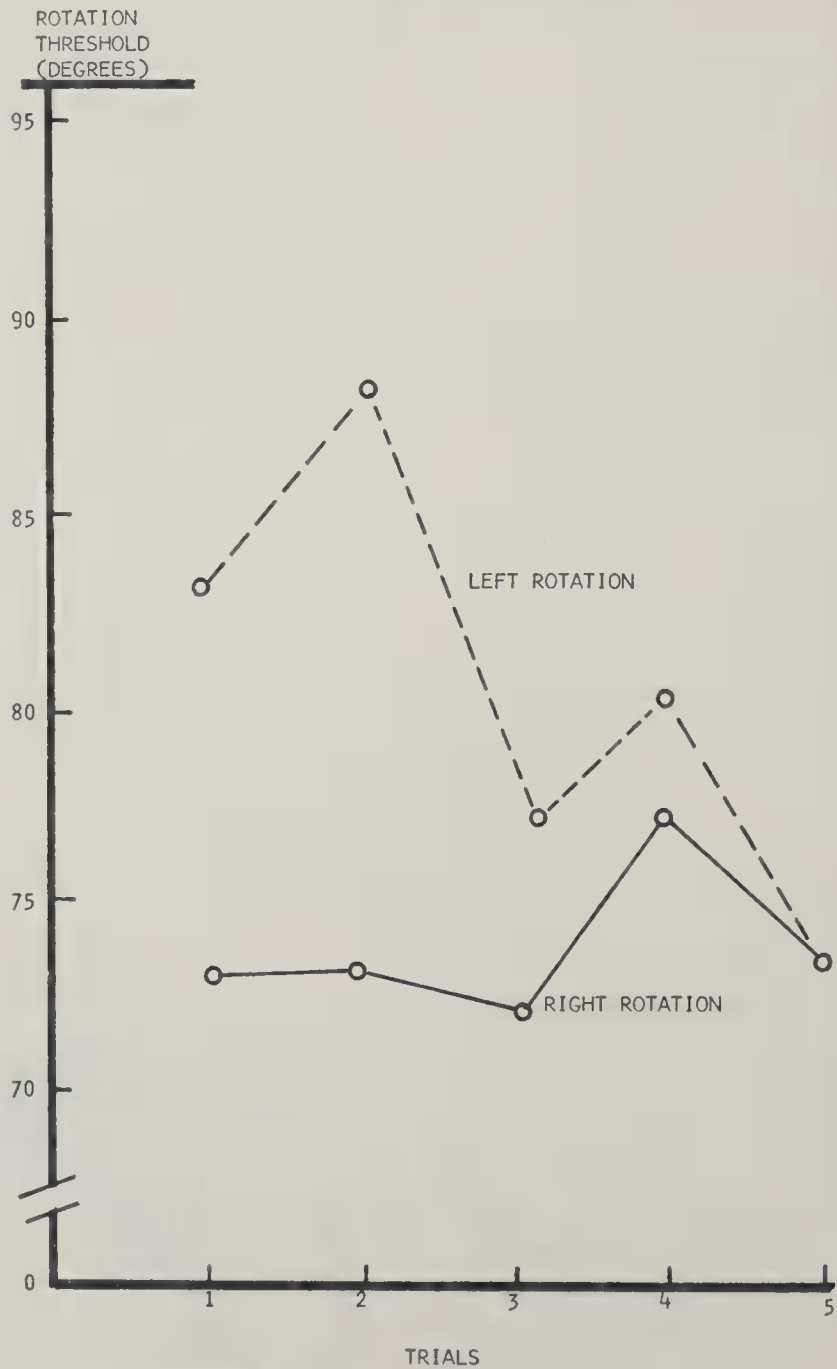


FIG. 3. The effects of practice on right and left rotational breakdown thresholds.

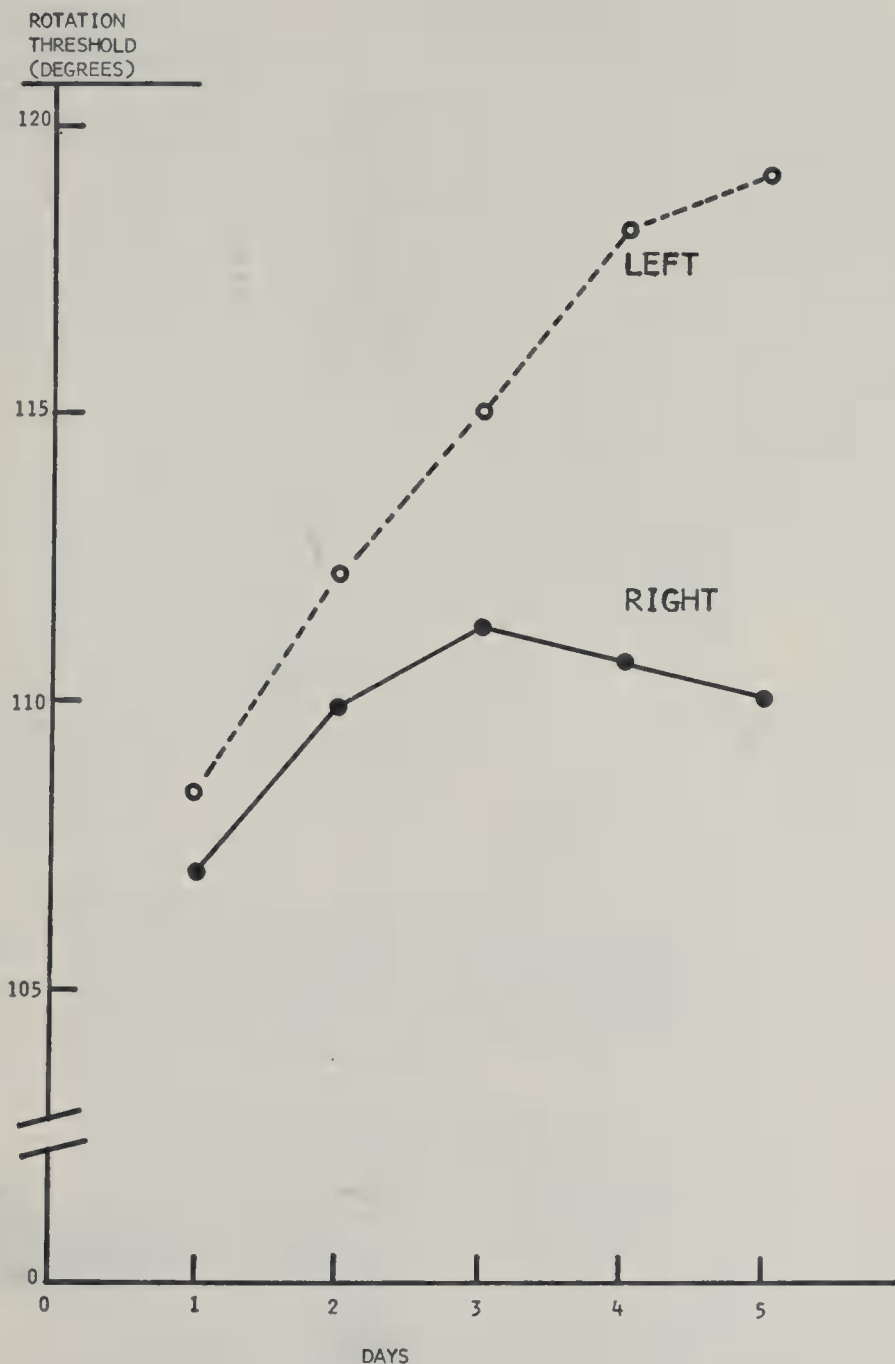


FIG. 4. Rotational breakdown thresholds to the left and to the right for left-handed Ss.

where *S* fell behind his normal pace, the motor rotating the system was stopped and the angle recorded. Although there are arbitrary judgments which must be made in establishing angular thresholds by this method, it was judged that the experiment would at least illustrate the possibility of obtaining differential rotational breakdown thresholds in reading.

Twenty-five university-student Ss were given 10 trials per day in reading, of which 5 were with left rotation and 5 with right rotation. The reading material used was a 3-inch column of *Reader's Digest*

material. Order of presentation of the right and left rotation was varied by Ss so that 13 started with rotation to the right and 12 with rotation to the left.

The data for this experiment are given in Figure 3. This graph plots rotation thresholds as a function of practice. As assumed, a definite rotational breakdown threshold was found for all Ss both to the right and to the left. The two curves compare the right and left rotational breakdown thresholds. Initially, those to the left are much larger than those to the right. However, this difference disappears with practice. The differences between the combined means for days

for the two conditions is statistically reliable at the 1% level, when the differences are tested by means of *F* tests.

Results

The results of this experiment generally confirm the assumptions that defined the design of the study. The rotational breakdown threshold for reading in right-handers is roughly 75 degrees. In the case of the right rotation, however, there is no real adaptation to this angular displacement. However, some adaptation occurs with left rotation, in which case the breakdown displacement angle is initially very much higher. From related observations, we know that the effects of these dynamic displacements on reading are much more severe than static rotation of single letters and words.

EXPERIMENT III: ROTATIONAL BREAKDOWN THRESHOLDS IN LEFT-HANDERS

Fifteen left-handed *Ss* were tested in a preliminary study for six trials per day for 5 days in order to determine their rotational breakdown thresholds. The method described above was used. Three of the daily tests were given with rotation to the left and three with rotation to the right. Order of presentation of conditions was varied with *Ss*. Three test trials were run for each *S* each day.

The results of this study are shown in Figure 4. These results indicate that *Ss* gave significantly higher rotational breakdown thresholds to the left, and the effect of practice on the left rotation was much more marked than that of right rotation. The breakdown thresholds found for these left-handers were considerably higher than those observed with right-handed *Ss* in the experiment described earlier. We do not know whether these differences would be found if *Ss* were run in the same experiments at the same time.

The statistical comparison between rotational breakdown thresholds for these left-handers is confounded by the fact that not all *Ss* show a larger breakdown threshold to the left than to the right. On three initial days of testing, six of the *Ss* used in the study showed larger rotational breakdown thresholds to the right. The overall difference between

left- and right-rotation thresholds, however, was statistically significant.

These observations on dynamic anisotropy in reading are the first to define a possible explicit relationship between handedness and reading and form perception. Although we consider the results purely of a preliminary nature they tend to suggest that eye-movement control and form perception are organized in terms of dynamic sensory-feedback integration of response in relation to the arrangement and orientation of forms in space.

EXPERIMENT IV: ANALYSIS OF SYSTEMATIC AND ANGULAR DISPLACEMENT OF READING FEEDBACK

A major parametric analysis of the anisotropic properties of reading behavior was done by testing the errors and time of reading for normally oriented, reversed, inverted, and inverted-reversed printed matter. A special reading display device was used for presenting the material at five different angles of static rotational displacement, i.e., 0 degrees, 45 degrees to the right, 45 degrees to the left, 90 degrees to the right, and 90 degrees to the left. Symmetrical displays of three-letter non-sense syllables were prepared and these were typed in four different displacement patterns, normally oriented, reversed, inverted, and inverted-reversed patterns.

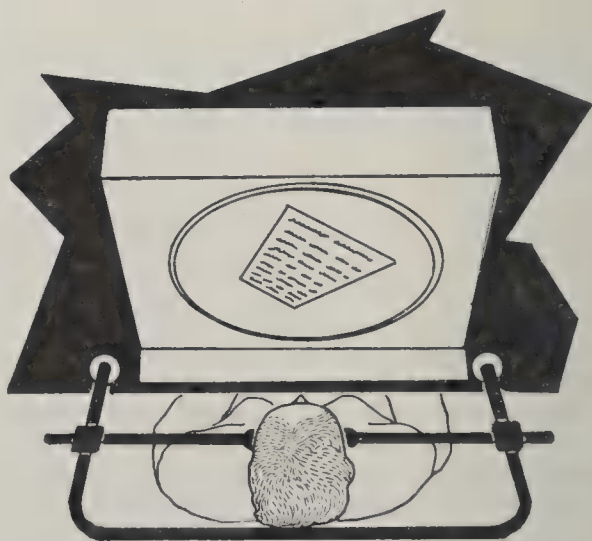


FIG. 5. Laboratory setup for measurement of combined systematic inversion and reversal and rotational displacement of sensory-feedback in reading.

Method

A diagram of the laboratory setup used is given in Figure 5. This illustrates the way in which S's head was positioned by means of the side supports of a headstand. The Ss were seated comfortably in front of the reading display unit, the angular position of which could be set at the values indicated.

The design of the experiment was based on adjustment of the order presentation of the angular displacement conditions to 12 Ss. Each S was run for approximately 1 hour each day at all five angular displacement conditions and with all four samples of syllables. Each S began the experiment by reading the normally oriented material at the zero displacement angle. Aside from this one exception, the order of presentation of the five angles of displacement and four orientations of printed matter was varied at random for 12 Ss. Twenty trials were thus presented to each S each day.

In running each trial for each S, a card containing the specially printed syllables was placed on the display apparatus at the proper angular displacement. The S then read aloud each nonsense syllable until he completed the card. Each of the trials was timed and errors recorded for checking the reading against correct copy.

Results

The main results of this experiment deal with the variation in errors and time of performance of the different orientations of printed matter at each of the angular displacement conditions. The results on errors are shown in Figure 6. The variations for general orientation of the printed matter are indicated in terms of differences in the blocks of bars that represent means of these conditions. The parameter within each of these blocks is displacement angle. The graph shows a marked variation in errors due to general directions of orientation of printed matter and a less marked variation due to angle of displacement. The inverted and reversed conditions give by far the most errors. The inverted-reversed condition shows a general level of error less than half that for the combined inverted-reversed orientations.

The main variation in reading error due to angle of displacement is related to the 45-degree left rotation. This angle produces

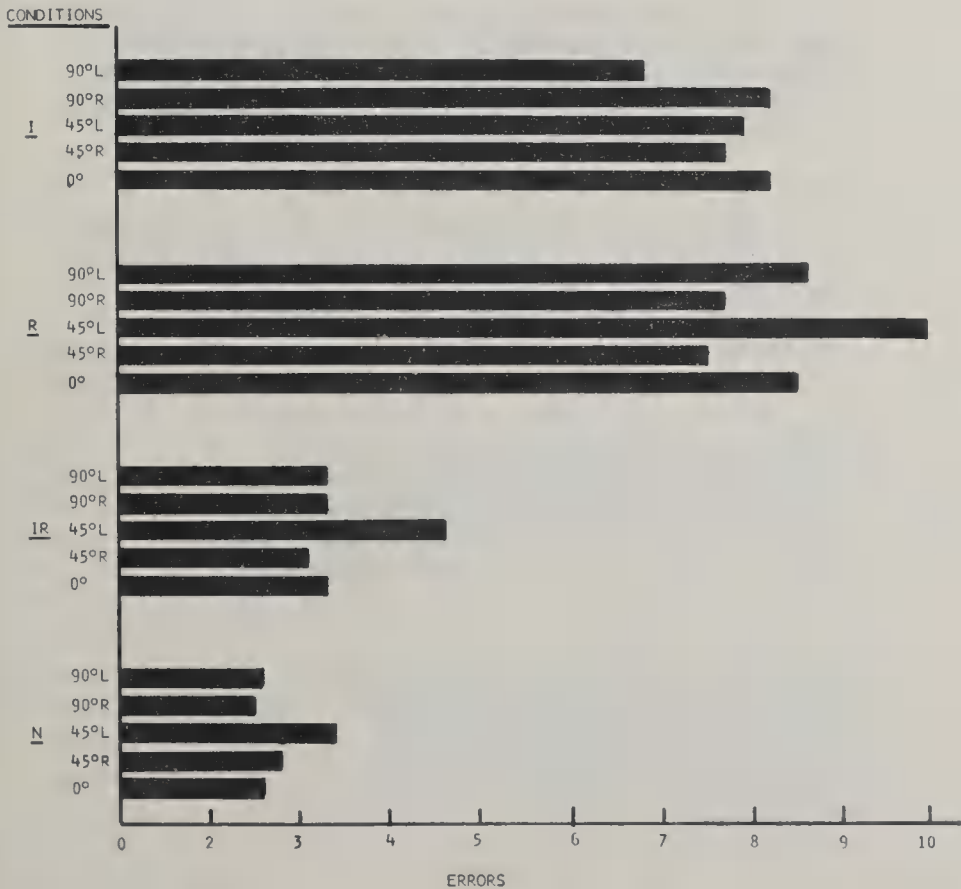


FIG. 6. Errors in reading at different rotational displacements for different parameters of systematic displacement of sensory feedback in reading.

a nearly significant level of error at all orientations of printed matter except the inverted condition. At this condition, all displacement angles produce much the same level of error except the position 90 degrees left. Overall, the results confirm the main assumptions that guided this research. That is, they indicate errors of reading are linked in differential interacting ways to both the general orientation of the printed matter and to angle of displacement.

Similar results on time of reading are shown in Figure 7. The order of effects for direction of orientation is much the same as that observed for error scores. However, the nature of the variation in time due to angles of displacement is much different from that observed with errors. The nature of the time variation due to angles of displacement is not statistically significant overall for angles of rotation but is significant at the 1% level for

conditions of orientation of the printed matter.

As found in the first study reported here, reversal tended to produce more errors than the inverted displacement, whereas the inverted condition caused a greater increase in reading rate than reversed feedback. For errors, this effect was observed at the zero, 45-degree left, and 90-degree left conditions, but was not observed at the other two angles of displacement. For the time measure, inversion was found to produce a greater time of reading at all angles of displacement. These differences, however, are not marked as they were in the first experiment, in which *S* groups always responded under one condition of systematic inversion or reversal.

A measure of overall efficiency in reading can be obtained for both time and errors combined by multiplying these two measures. The results on this index of efficiency at different

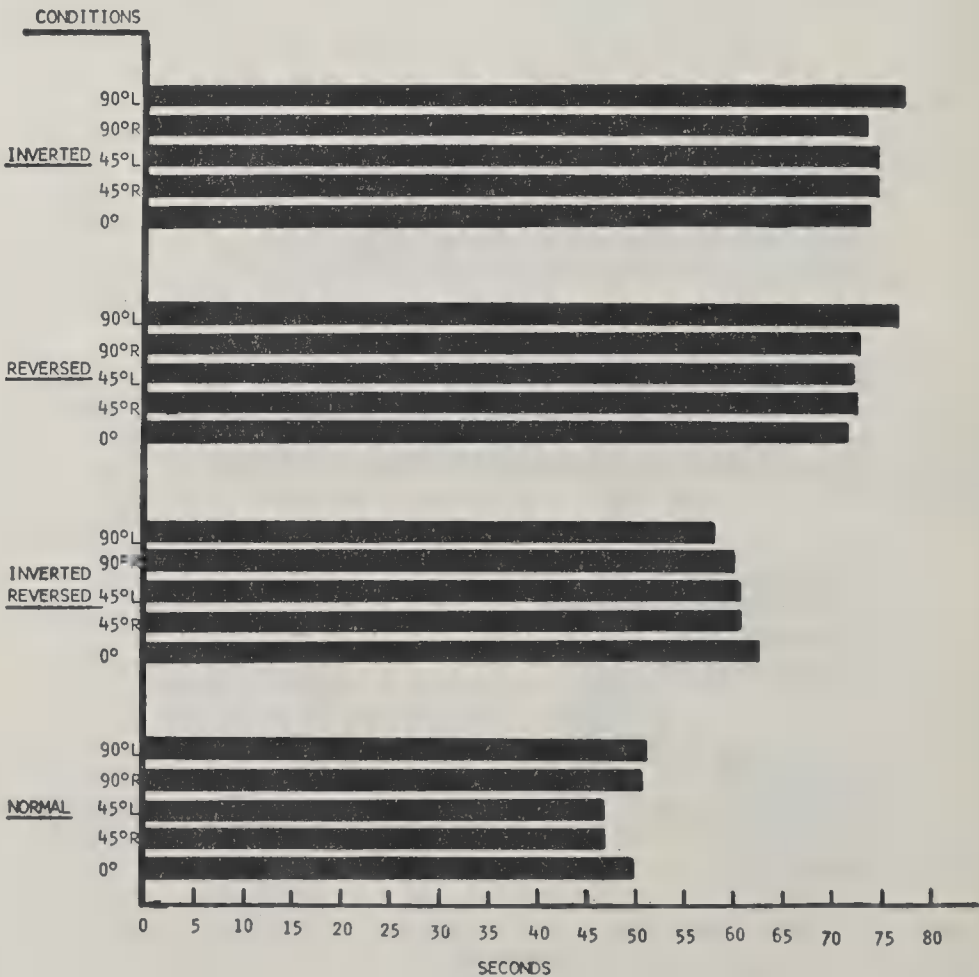


FIG. 7. Rate of reading at different rotational displacements for different parameters of systematic displacement of visual feedback in reading.

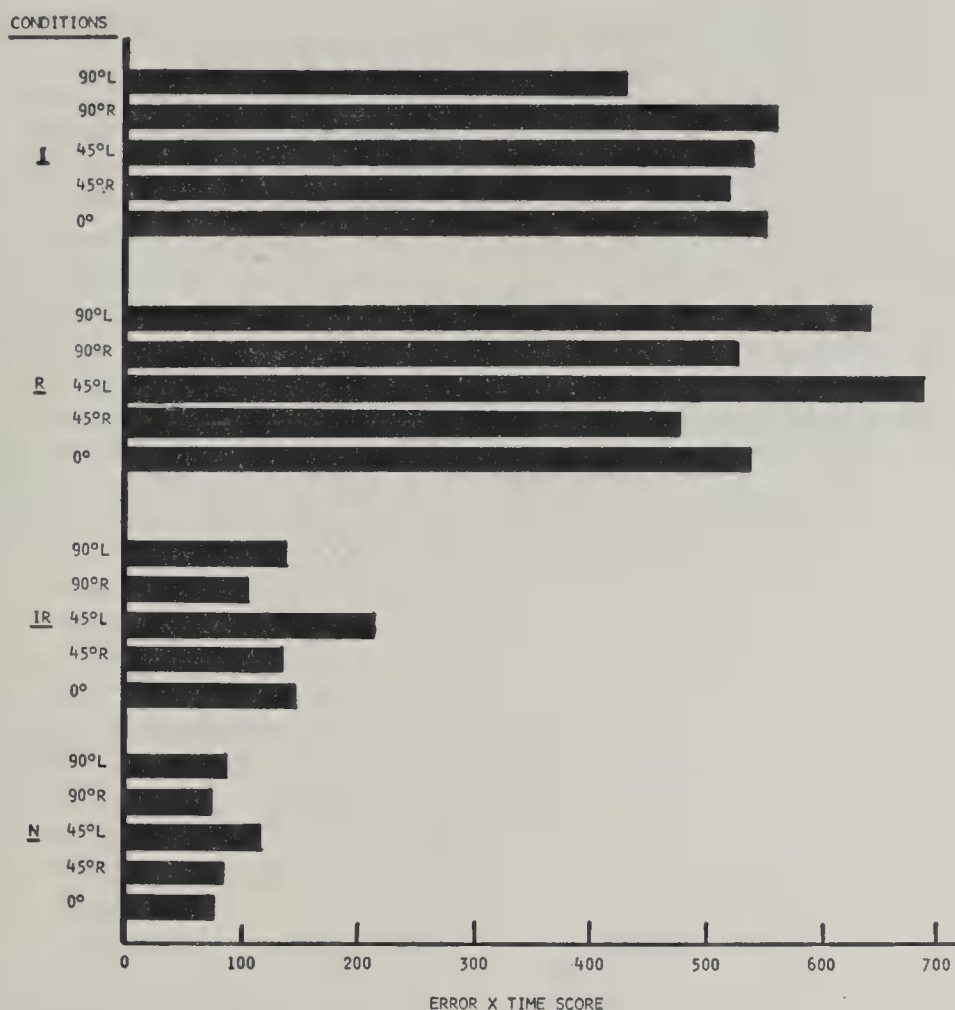


FIG. 8. Combined error and time scores at different rotational displacements for different parameters of systematic displacement of visual feedback in reading.

conditions of orientation and angles of displacement are given in Figure 8. As expected, this graph reflects what we have said before about errors and angles of displacement since the variation in time of reading relative to this parameter is limited. The overall efficiency of reversed reading is poorest with the left rotational positions of 90 and 45 degrees, whereas the overall efficiency of inverted reading is poorer at all angles but 90 degrees left. In the case of inverted-reversed reading, efficiency is poorest at 45 degrees left and about the same at the other four displacements. This same pattern is found for normally oriented reading.

SUMMARY AND DISCUSSION

These experiments have been designed to use simple and direct materials and procedures in order to demonstrate the application of

space-displaced sensory-feedback methods to the study of reading. The main results found were these:

An interesting and apparently significant area of research on reading behavior can be based on its differential anisotropic properties relative to the manifold conditions of orientation of reading material in space.

Results show that inversion and reversal of printed matter have very similar effects on the pattern and properties of behavior in reading. Combined inversion and reversal had far less effect than either of these two unidimensional directions of displacement. Under inversion and reversal, errors in reading were increased by 400 to 500%, while with combined inversion and reversal they were not quite doubled.

In keeping with assumptions, a rotational breakdown threshold for reading was found

in all of the right-handed Ss observed. In addition, this threshold displayed differential magnitudes for right and left rotation of reading matter in these right-handed Ss. The magnitude of these thresholds was approximately 75 degrees after four periods of practice. The practice effect appeared only for left rotation. The rotational breakdown thresholds were fairly consistent in individual Ss and reproducible from day to day. Although the methods tried out here have certain defects in establishing exact rotational thresholds for reading, they have been sufficiently successful to suggest that precision technical diagnostic reading instruments can be developed for such rotational threshold measurement.

Preliminary observations of the magnitude of rotational breakdown thresholds in left-handed Ss generally conform to assumptions about the relationships between handedness and response to dynamically displaced sensory data in reading. The left-handers tested give larger rotational breakdown thresholds to the left. In addition, the magnitudes of their thresholds both to the right and left were greater than for right-handers. Because there are certain arbitrary judgments which must be made in judging angular displacement breakdown thresholds by means of this method further observations need to be done to compare such thresholds in right- and left-handers.

The effects of rotational displacement in reading statically positioned material is not as marked as the events observed with dynamic displacement. Displacement angles of 45 and 90 degrees showed variations in errors of reading that approached significance with particular types of systematic inversion and reversal of the printed matter. There are definite patterns of interaction between the general direction of spatial orientation of printed matter and statically displaced angular rotation of sensory-feedback in reading.

The results on rotational breakdown thresholds and on inversion and reversal of reading specifically confirm our view that the so-called perceptual processes in display behavior are dynamically organized in terms of generation of geometric visual data by movements. Both time in reading and the time needed to recognize displays with specific spatial orientation

are determined by the anisotropic specialization of sensory-feedback effects of perceptual behavior. Not only does the appearance of print change with orientation in space, but the characteristics of dynamic behavior essential for reading also vary in ways that seem to have a certain uniformity. The data from the inversion and reversal study suggest that, like other human motions, reading movements are organized in terms of two main intrinsic reference axes of sensory-feedback control of motion, i.e., the built-in reference mechanisms related to the gravitational control of posture (which references the vertical axis of motion) and the bilateral transport movement system (which references the horizontal axis of motions). However, unlike the study of the control of manual motions, the present study has found no marked difference between inverted and reversed conditions of feedback upon reading performance. Instead, the main differences have been found between these two conditions of feedback and the combined inverted-reversed condition. There is little doubt that the specific pattern of certain letters, as evolved in the evolution of writing, determines the results that have been found in these studies.

The results indicate that the anisotropy of display behavior can be used to devise new ways to explore individual differences in reading, reading disability, role of handedness and eyedness factors in reading, display learning, and to devise preapprenticeship reading training devices that may aid beginning reading in the child. In the terms defined here we can begin to study and diagnostically analyze reading and form perception in terms of its true status as a dynamic psychomotor feedback process and begin to examine critically the reinforcement concepts (Mowrer, 1963) in relation to capability for actual understanding of educational skills.

The type of experimental analysis that we have done here on reading is not a denial of the role of the learning factor in perception, but an effort to probe experimentally into the stimulus-response process itself, as a means of understanding learning. Thus, the present experiments extend direct analytic behavioral methodology to the perceptual sector. The findings suggest that there are complex space-organized dynamic behavioral activities of the

eye-movement system and of postural and transport behavior whose development and integration as sensory-feedback mechanisms may precede the learning of reading. Thus our experiment may define a general psychomotor preapprenticeship condition for learning of successful reading that emphasizes training the child in diversified conditions of space-displaced vision of his own body and manual movements.

The theory of the experiments and the results contain some suggestive ideas about the future study of handedness in relation to reading and other speech disabilities. If the basic mechanism of organization of motion consists of built-in inertial reference mechanisms of sensory-feedback control of movements in main axes of the body, as the present theory suggests, then handedness may be an expression of genetic reversal of the bilateral reference system for all movements in the horizontal axes of motion, and not just a directional reversal of specific motor mechanisms. Such a disturbed reference mechanism could also affect the motions of speech, reading, and writing during maturation. Future studies can be done to test specific hypotheses of this nature by means of sensory-feedback analysis of reading, handwriting, drawing, and speech. The finding that static displaced reading of single characters may not be related to handedness in very young children (Flescher, 1962) should not be interpreted too

rigorously because in most children the mechanisms of handedness may not be firmly organized until the child is 11 or 12 years old (Greene & Smith, 1963).

The present experiments also can be related to a basic hypothesis of the theory (Smith & Smith, 1962) that sensory-feedback mechanisms of behavior are differentially organized so that inversion of feedback in the vertical axis of the body will be different from reversal from right to left. This assumption is related to the theory that the postural-gravitational system and the bilateral-transport mechanisms of response not only regulate their respective movements directly, but also act as intrinsic or built-in inertial reference systems for all manipulative and finer motions. The prediction in this research is that reading behavior is organized spatially as a sensory-feedback process in relation to these postural and bilateral movement reference systems, and that differential effects of inversion, reversal, and angular rotation of printed matter will be found for reading activity.

Figure 9 gives a general idea of this neurogeometric concept of the brain's detection of movement-generated spatial data in reading and form perception. This diagram shows that every eye or head movement in reading is controlled by direction-specific neurons lodged in the brain and that the geometricity of these neuron systems determines the anisotropic properties of reading behavior. The origin of

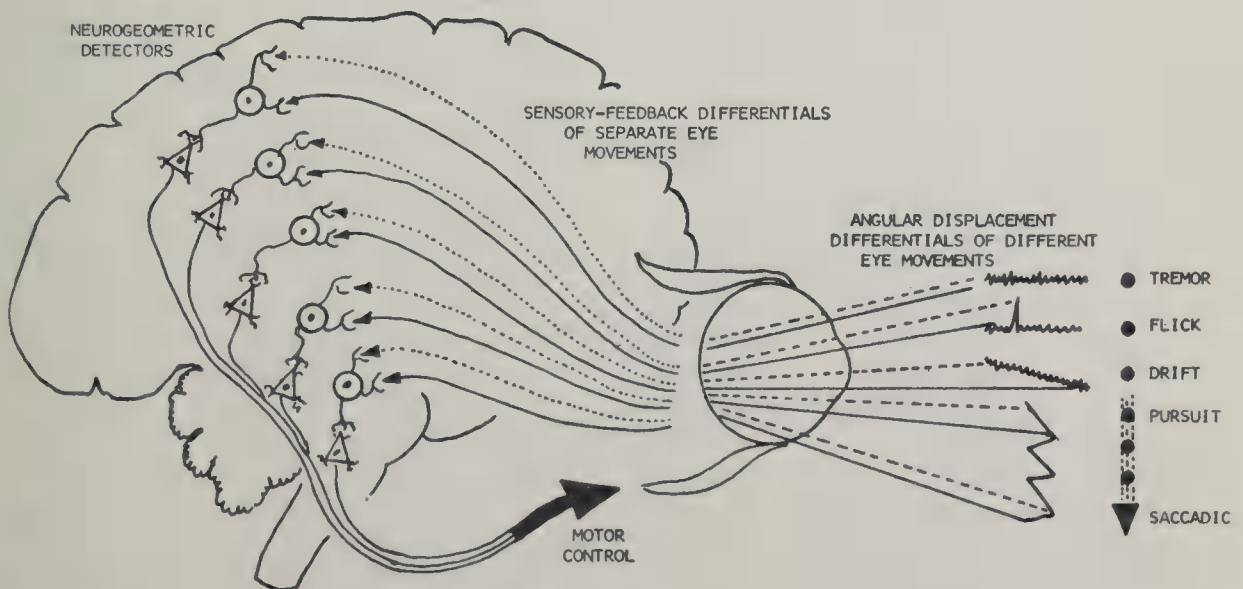


FIG. 9. The neurogeometric concept of control of the basic patterns of eye movement, including fixation and pursuit movements of reading.

the belief in such direction-specific cells in the visual system lies in our original studies of the directionally specialized neural mechanisms of optic nystagmus in animals (Smith & Bridgman, 1943). The view derived from these studies was that each internuncial neuron of the visual system acts as a space recorder to detect stimulus differences between a given reference system of a movement and its movement-generated visual feedback. That is, every component of eye movement—fixation, tremor, fixation flick, fixation drift, pursuit, saccadic, and compound patterns of binocular movement accommodation—is independently regulated by separate ganglionic detector systems that record differences in movement-generated spatial visual data and direct the eye in differential ways in reading in accordance with the magnitude and direction of the sensory-feedback related to the movements. The recent work of Barlow and Hill (1963) on the rabbit's eye and of Hubel and Wiesel (1959, 1962) on the cat's cortex are interpreted as supporting this set of ideas. These investigators have found direction-specific ganglia in the retina and in the cortex of the visual system of animals.

1. This research applied methods of space-displaced sensory-feedback analysis to the investigation of reading in order to test whether anisotropic properties of dynamic behavior define the nature of organization of this verbal activity.

2. Results show that different conditions of systematic displacement of visual feedback of printed matter produce differential effects on the error and rate of reading.

3. As specified by assumptions, limited ranges of normal and breakdown angular displacement of the visual feedback of reading are found with rotation of printed matter.

4. The rotational breakdown thresholds of reading vary with direction of the displacement.

5. The magnitude of rotational breakdown thresholds to the right and to the left differs in right- and left-handed Ss. The results are the first clear-cut behavioral data that relate reading behavior to handedness.

6. The findings confirm the main hypothesis that the anisotropic properties in reading

and form perception are related to spatial organization of the sensory-feedback mechanisms of eye-movement responses and to motion reference systems that regulate the direction of these movements in space.

7. The interaction between the systematic orientation of printed matter and the effects of angular displacement suggest that the figure-ground process in reading, form perception, and other attentional forms of behavior is a direct product of the interaction between the built-in sensory-feedback reference systems of response and the relative angular displacement of exteroceptive stimuli of focal movements in the response pattern.

8. The experiments' methods and results are discussed in relation to their possible future implications for the theory of reading and form perception, and especially in defining stimulus-response techniques of analysis of the psychophysiology of perceptual organization in reading and shape perception.

REFERENCES

- BARLOW, H. B., & HILL, R. M. Selective sensitivity to direction of movement in ganglion cells of the rabbit retina. *Science*, 1963, **139**, 412-414.
- FLESCHER, I. Ocular-manual laterality and perceptual rotation of literal symbols. *Genet. Psychol. Monogr.*, 1962, **66**, 3-48.
- GREENE, P., & SMITH, K. U. A critical period in maturation of performance with space-displaced vision. *Science*, 1963, **141**, 727-728.
- HUBEL, D. H., & WIESEL, T. N. Receptive fields of single neurons in the cat's striate cortex. *J. Physiol.*, 1959, **148**, 574-591.
- HUBEL, D. H., & WIESEL, T. N. Receptive fields, binocular interaction, and functional architecture in the cat's visual cortex. *J. Physiol.*, 1962, **160**, 106-154.
- MACH, E. *Analysis of sensations*. (Trans. by C. M. Williams) Chicago: Open Court, 1897.
- MOWRER, O. H. Learning theory and pedagogical practice. In V. Herrick (Ed.), *New horizons for research on handwriting*. Madison, Wis.: Univer. Wisconsin Press, 1963.
- RHULE, W., & SMITH, K. U. Effect of visual pre-training in inverted reading on perceptual-motor performance in inverted visual fields. *Percept. mot. Skills*, 1959, **9**, 327-331.
- SMITH, K. U., & BRIDGMAN, M. The neural mechanisms of movement vision and optic nystagmus. *J. exp. Psychol.*, 1943, **33**, 165-187.
- SMITH, K. U., & SMITH, W. M. *Perception and motion: An analysis of space-structured behavior*. Philadelphia: Saunders, 1962.

(Received August 29, 1963)

THE ATTITUDES OF RESEARCH CHEMISTS

JOHN R. HINRICHS¹

Data Processing Division, International Business Machines Corporation, White Plains, New York

Principal component analysis of questionnaire data from roughly $\frac{1}{3}$ of the nation's PhD graduates in chemistry in 1960-1961 isolated 3 basic attitude patterns: (a) attitudes valuing freedom and "pure science," (b) materialistic attitudes accepting business values, possibly at the expense of science values, and (c) attitudes which see little conflict between industry and science values. New PhD's with high pure science attitudes tended to enter academic employment; others to enter industry. For an independent sample of 286 industrial chemists, both the orientation to "applied science" and the materialistic orientation were stronger for chemists with high number of years experience than for recent hires.

"In the 1960's, our economy has firmly turned the corner from the industrial revolution to an era of science"—this is a theme which is often voiced among observers of our contemporary scene. At the same time, it is clear, we have now entered an era of institutionalized science. No longer is the "scientist" stereotyped as a lonely garret inventor. Instead, we more frequently think of "task forces" and teams of scientists employed by bureaucracies—government laboratories, research foundations, university research centers, or industrial laboratories. Inevitably, certain marked strains have arisen as a result of these changes in the traditional institutions of science.

The literature dealing with the administration of scientific research usually attributes the major strains to conflicts between professional identifications and institutional identifications, particularly in the case of industrial scientists (Danielson, 1960; Marcson, 1960; Pelz, Mellinger, & Davis, 1954). This conflict has been cited as the cause of relatively low morale among industrial research personnel (Moore & Renck, 1955; Opinion Research Corporation, 1959).

Most current literature on research management proposes administrative action designed to reduce this conflict. The general

flavor suggests a relatively "open system" (Barnes, 1957) in which concern for the individual scientist tempers management concern for the goals of productivity and practicality. An authority system based on the relationship between senior and junior professional colleagues is recommended instead of the traditional bureaucratic boss-subordinate relationship (Brown, 1957). At the very least, a participatory pattern of supervision is advocated for the research supervisor (Baumgartel, 1956).

In spite of recommendations that the organization adapt itself to accommodate its research personnel, however, most adjustments to alleviate the stresses associated with the institutionalization of science probably consist of changes on the part of individual scientists; either changes in jobs or modification in some degree of their value systems in order to fit in with the industrial environment.

To provide further data on the interaction between a scientist and the organization which supports him, the research reported in this article deals with four aspects of the profession-oriented attitudes of research chemists: (a) It investigates some aspects of the nature of individual differences in science-related attitudes and identifications; (b) It evaluates some of the factors which influence a chemist's choice of a career at the time he leaves the graduate school setting; (c) It assesses the impact of exposure to the industrial environment on the attitudes of research chemists; and (d) It investigates the relationship between patterns of profession-

¹This study is based on a dissertation submitted to the Graduate School of Cornell University in partial fulfillment of the requirements for the PhD degree. The author would like to express his appreciation for the advice and assistance given him by F. F. Foltman, H. A. Landsberger, P. J. McCarthy, and P. C. Smith of Cornell University.

oriented attitudes and work satisfactions for chemists employed within the industrial environment.

METHOD

Because they represent a large population of scientists and because their work is relatively similar across a number of environments, chemists who had completed their PhD requirements were selected as the sample for the study.

Questionnaire Development

A pool of over 100 Likert-type statements was developed from literature dealing with the administration of scientific research and from interviews with industrial chemists. Statements dealt with a wide range of attitudes toward science, factors affecting occupational choice, and the perceived role of the individual within the chemistry profession. Half were phrased in favorable terms and half in unfavorable terms.

This preliminary questionnaire was administered to all graduates in chemistry at eight East Coast universities who had either already received the PhD degree during the previous 6 months or who were scheduled to attain it within the next 6

months. Thus, the data covered graduates during a 1-year period and sampled their attitudes as near as possible to the time that they attained the degree.

One hundred and fifty-six chemists were surveyed—some following a personal interview at their school and some by mail. Based on replies by 76% (115 people), ambiguous items and items with highly skewed responses were deleted. The final refined questionnaire consisted of 79 Likert-type items, questions dealing with job satisfaction, and personal history items.

Sampling

Two samples were used:

1. One sample, which consisted of roughly $\frac{1}{3}$ of the population, was designed to represent all new graduates in chemistry in the United States for the 1-year period between February 1960 and February 1961. First, 41 of the 96 United States universities granting the PhD degree in chemistry were selected based on a random sampling design stratified in terms of size of university. Next, the questionnaire was mailed to 491 graduates of these schools selected under a probability sampling design. Replies were received from 385, or 77%. There were no significant differences in response rates between large and small universities.

TABLE 1
ITEMS SCORED FOR COMPONENT I

A lot of times in industry a good research chemist is stifled by being put under a supervisor who is not technically qualified.	.24
Most recognition for industrial research goes to management or the company rather than to the individuals who did the work.	.23
In industry, scientists' talents are channeled too closely to what is proven and profitable.	.23
Many good scientific ideas have "died on the vine" because they have not received adequate support from people in authority.	.22
Company security regulations keep a lot of good industrial research results out of literature.	.21
Most companies look on scientific research solely as a means of protecting profits by keeping ahead of competition.	.21
Getting ahead in industry is based more on politics than it is on knowledge.	.20
There are too many low-grade "hacks" doing chemical research today.	.20
To encourage effective research, industrial laboratories should try to create an atmosphere similar to a university setting.	.19
A chemist must have freedom to formulate and develop his own research ideas to produce significant research results.	.18
Too many industrial jobs require a man to move to a different location every so often.	.17
Social conformity and scientific creativity are incompatible.	.17
In general, chemists going into industry don't have any trouble making their scientific goals jibe with the company's objectives.	— .16
The only thing holding back a surge of creative research is insufficient financial backing.	.16
It's a waste of talent to use chemists who were in the upper half of their college graduating class in administrative jobs.	.16
In any organization, the people in power get there by manipulating other people.	.15

TABLE 2
ITEMS SCORED FOR COMPONENT II

Most scientists are more interested in their profession than in an opportunity to move up in the organization for which they work.	-.22
A chemist can put up with monotonous work if the pay is ok.	.20
To any chemist "worth his salt," the most important thing in a job is the opportunity to do sound scientific work.	-.19
To encourage effective research, industrial laboratories should try to create an atmosphere similar to a university setting.	-.18
Adequate financial rewards are of major importance in getting scientists to do the best possible job.	.18
In any organization, the people in power get there by manipulating other people.	.18
A lot of research people elaborate and add more detail to a job than is needed.	.17
Many chemists waste time trying to investigate every possible angle before making a decision.	.17
A chemist must have freedom in applying his own ideas to solve technical problems if he is to produce significant research results.	-.17
Everyone wants a chance to gain a certain amount of power over other people.	.17
A scientist is bound to have a somewhat limited perspective unless he works on field problems and applications at least once in a while.	.17
A chemist must have freedom to formulate and develop his own research ideas to produce significant research results.	-.16
Scientific research is more meaningful if the researcher has a chance to see how well his findings work in an applied situation.	.16
In a university setting there is a considerable amount of pressure to conform to established methods of solving problems.	.16
Scientists who let an organization force them to do a slipshod research job are really dishonest to the profession.	-.16
Supervisors of industrial research don't necessarily have to be top-notch technical men.	.16
The primary goal of scientific research is to raise the standard of living in this country.	.16
A lot of times in industry a good research chemist is stifled by being put under a supervisor who is not technically qualified.	.16
Women think it is more glamorous for a husband to be an industrial executive than a research chemist.	.15
Many good scientific ideas have "died on the vine" because they have not received adequate support from people in authority.	.15
If a man's salary is high enough, he "feels" recognized.	.15

2. A second sample represented three different industrial organizations—a pharmaceutical laboratory, a petroleum research laboratory, and a general chemical company. Three hundred and seventy-four PhD level research chemists were surveyed; 286 responded (76%). There were no significant differences in response rates among the three companies.

ANALYSIS AND RESULTS

Data for these two samples, the "new graduates sample" and the "industrial sample," were analyzed separately.

New Graduates Sample

A principal component analysis (Hotelling, 1933) of item intercorrelations (Pearson r) was intended as a statistical technique to summarize 64 selected questionnaire items into a relatively small number of linearly independent variates. The first three components extracted clearly differed from one another and accounted for 7.8%, 5.6%, and 4.3%, respectively, of the total variability among the 64 items. The percentage of variance accounted for by subsequent components

TABLE 3
ITEMS SCORED FOR COMPONENT III

Scientific research is more meaningful if the researcher has a chance to see how well his findings work in an applied situation.	.32
It's important to belong to an organization that is well thought of in the community.	.27
The research man who has a practical outlook on problems gets the best results.	.25
A scientist is bound to have a somewhat limited perspective unless he works on field problems and applications at least once in a while.	.23
Most chemists are interested in pursuing knowledge for its own sake.	.22
In general, chemists going into industry don't have any trouble making their scientific goals jibe with the company's objectives.	.21
High professional standards of chemical research must be maintained at all costs.	.21
The primary goal of scientific research is to raise the standard of living in this country.	.19
A scientist going into industry owes his first loyalty to his company.	.18
The only thing holding back a surge of creative research is insufficient financial backing.	.18
Getting ahead in any technical job depends primarily on a man's technical competence.	.18
To any chemist "worth his salt," the most important thing in a job is the opportunity to do sound scientific work.	.18
A chemist must have freedom in applying his own ideas to solve technical problems if he is to produce significant research results.	.17
Most significant basic chemical research is done in universities.	-.17
A chemist can put up with monotonous work if the pay is ok.	-.17
As a rule, research chemists in industry are able to work on projects which make full use of their individual training and skills.	.16
Scientists who let an organization force them to do a slipshod research job are really dishonest to the profession.	.16
Getting ahead in industry is based more on politics than it is on knowledge.	-.15

tended to level off, and it was not possible to develop a reasonable explanation of the attitude areas being tapped. As plots of the first three components indicated that rotation of the reference axes would do little to improve the interpretation given to them, unrotated solutions were used.

Questionnaire Scoring. For items with loadings greater than .15, raw-score responses (based on the scale, 1 = "agree" to 5 = "disagree") were weighted by the coefficient values of the characteristic vectors and summed as scores for each individual on each of the three components. Since all of the original variates had standard deviations very nearly equal to one it was not considered necessary to adjust them to have unit standard deviation before weighting. The items scored for each component with their loadings are shown in Tables 1, 2, and 3.

Low numerical scores on Component I appeared to reflect attitudes valuing freedom and support in research and a belief that industry raises barriers to worthwhile scientific activity. A low score resulted from responses generally agreeing with the items in the component.

Low scores on Component II apparently reflect relatively expedient attitudes and acceptances of business values, possibly at the expense of science values.

Agreement with the items scored for Component III was interpreted as indicating acceptance of industrial research with the belief that there is little serious conflict between science and industrial values.

Means and standard deviations for scores on these components are shown in Table 4.

Occupational Choice. It was hypothesized that new graduates in chemistry who had

decided on an academic career, when compared with chemists who had decided to enter industrial research, would: (a) score lower (tend to "agree") on Component I, (b) score higher ("disagree") on Component II, and (c) score higher on Component III.

To test these hypotheses, the sample of new graduates was divided into two groups: 222 chemists who planned a career in industry, and 152 people who had accepted academic employment, planned to, or who were currently taking postdoctoral work. A few individuals had not specified their occupational plans and were omitted from this phase of the study.

As shown in Table 5, these three hypotheses were supported by the data. However, the differences were small and there is limited practical significance permitting use of these scores to predict occupational choice of chem-

TABLE 4

MEANS AND STANDARD DEVIATIONS FOR SCORES ON THE THREE COMPONENTS FOR 374 NEW GRADUATES

	<i>M</i>	<i>SD</i>
Component I	7.53	1.36
Component II	5.07	1.22
Component III	7.61	1.24

ists. A discriminant function analysis yielded weak though significant discrimination between academic and industrial bound chemists based on these scores.

Industrial Sample

Effects of Time in Industry. It was hypothesized that chemists with relatively little experience in industry, when compared to chemists with high years of experience, would tend to: (a) "agree" with the items scored for Component I, (b) "disagree" with the items scored for Component II, and (c) "disagree" with the items scored for Component III. To evaluate these effects, the sample of 286 industrial chemists was divided into 10 groups of approximately the same size based on the number of years of full-time experience in chemistry which they reported. Their questionnaires were scored using the

TABLE 5

COMPARISON OF MEAN COMPONENT SCORES: NEW GRADUATES ENTERING INDUSTRY VERSUS THOSE ENTERING ACADEMIC WORK

	Academic bound ^a		Industry bound ^b		<i>CR</i>
	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	
Component I	7.36	1.13	7.64	1.50	2.04
Component II	5.25	1.17	4.95	1.25	2.35
Component III	7.91	1.25	7.41	1.21	3.87

^a *N* = 152.

^b *N* = 222.

weights developed from the new graduate data. Mean scores on each of the three Components for each of the 10 "years of experience" groups are plotted in Figure 1.

Analysis of variance (Table 6) indicated that industrial chemists with high years of experience had significantly higher scores on Components II and III than did relatively inexperienced chemists. Results were not statistically significant for Component I scores.

Job Satisfaction. Responses to the following three questions (which could be answered on a five-point satisfaction scale) were summed for each industrial chemist as a general "satis-

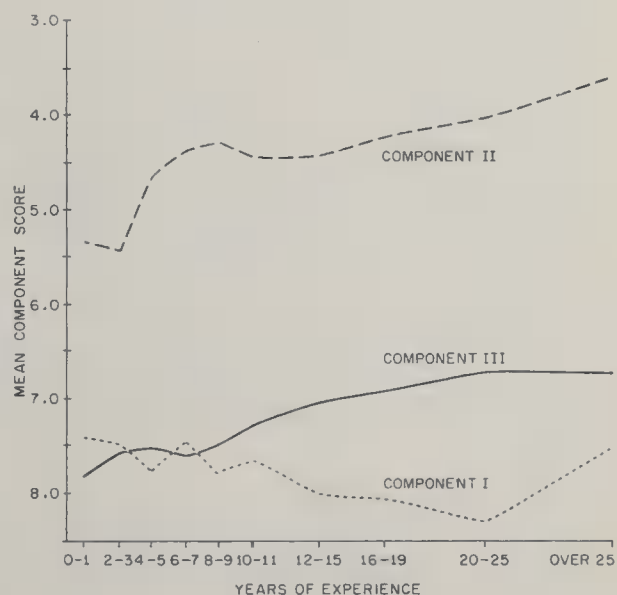


FIG. 1. Mean component scores versus strata for "Years of Experience." (Low numerical scores indicate high "agree" responses on the items scored for the component.)

TABLE 6

ANALYSIS OF VARIANCE TABLES FOR STRATA FOR
"YEARS OF EXPERIENCE," INDUSTRIAL SAMPLE

Source	SS	df	MS	F
Component I				
Between experience strata	24.21	9	2.69	1.46
Error	509.98	276	1.85	
Total	534.19	285		
Component II				
Between experience strata	75.38	9	8.38	5.50*
Error	420.45	276	1.52	
Total	495.83	285		
Component III				
Between experience strata	38.57	9	4.28	2.59*
Error	456.57	276	1.65	
Total	495.14	285		

* $p \leq .01$.

faction score": (a) "How satisfied are you with the latitude you have to attack problems in your own way?" (b) "How satisfied are you with the opportunities you have to do creative work?" (c) "How satisfied are you with the consideration given to your ideas?" High satisfaction scores were negatively related to Components I and II, and positively related to Component III (Table 7). Thus, high work satisfaction in industrial research appears to go along with acceptance of the value system of science while at the same time accepting the industrial system. The results suggest that work satisfaction tends to be relatively low in cases in which either of these patterns is held alone; i.e., a single-minded strong basic science orientation or a single-minded strong commercial orientation.

DISCUSSION AND IMPLICATIONS

It may be an oversimplification to speak solely in terms of "science orientation" versus

TABLE 7

CORRELATIONS BETWEEN THREE-ITEM SATISFACTION
SCORES AND EACH OF THE THREE COMPONENTS

Component I	$r = -.32$
Component II	$r = -.21$
Component III	$r = +.18$

"institutional orientation" as developed in other studies dealing with motivation in research. Perhaps the science orientation dimension should be subdivided in terms roughly analogous to a "pure research—applied research" division. These attitude dimensions are measurable. They also serve to some extent as determiners of occupational choice; new graduates tend to enter the occupational environment which fits their individual pattern of attitudes. Also, time apparently continues the process of fitting industry's scientists to the industrial mold.

Acculturation within industry is least evident for those attitudes valuing freedom and unimpeded scientific activity (Component I). However, the negative relationship between this Component and satisfaction scores suggests that holding to the "pure science" orientation within industry carries the price of relative work dissatisfaction.

Relative work dissatisfaction also tends to accompany the pattern of attitudes reflecting little science dedication and a materialistic and expedient philosophy (Component II). The strength of this pattern increases relatively strongly with increasing time in industry.

The most happy blend of attitudes appears to be for those chemists who are oriented toward science but who also see little conflict within the industrial research environment (agree with Component III). This pattern increases with time in industry and also is accompanied by relatively high satisfaction with industrial employment.

The surface implications are that the industrial research organization should beam its recruitment efforts to people with the Component III pattern of values. It should also, to the extent possible, structure its interaction with its scientists to channel the ongoing process of acculturation to encourage these attitude patterns. On an operational level, this means emphasizing the fact that research can be challenging, diverse, interesting, worthwhile, and of general importance, even if applied, rather than contributing to the misapprehension that the "true scientist" is interested only in fundamental research.

Just as it is possibly a mistake for industry to play down the fact that most of its research is applications oriented, the results tend to reinforce Herzberg's (Herzberg, Mausner, & Snyderman, 1959) conclusions that it is probably also a mistake for industry to emphasize the material aspects associated with a career in industrial research: salary, status, power, etc. The Component II attitude pattern which focuses on such elements is not associated with science identification and is not accompanied by high work satisfaction.

In drawing implications from this study, it is evident that a major element is missing: research performance. Unfortunately, reliable performance data were not available. Rough indications of performance were supplied by the cooperating companies, but these were largely a reflection of salary levels and not sufficiently comparable between companies for use as performance criteria.

REFERENCES

- BARNES, L. B. *Organizational systems and engineering groups, a comparative study of two technical groups in industry*. Boston: Harvard University, 1957.
- BAUMGARTEL, H. Leadership, motivations, and attitudes in research laboratories. *J. soc. Issues*, 1956, 12, 24-31.
- BROWN, PAULA. Bureaucracy in a government laboratory. In R. T. Livingston & S. H. Milberg (Eds.), *Human relations in industrial research management*. New York: Columbia University Press, 1957.
- DANIELSON, L. E. *Characteristics of scientists and engineers significant for their motivation and utilization*. Ann Arbor: University of Michigan, 1960.
- HERZBERG, F., MAUSNER, B., & SNYDERMAN, BARBARA B. *The motivation to work*. New York: Wiley, 1959.
- HOTELLING, H. Analysis of a complex of statistical variables into principal components. *J. educ. Psychol.*, 1933, 24, 417-520.
- MARCSON, S. *The scientist in American industry*. New York: Harper, 1960.
- MOORE, D. G., & RENCK, R. The professional employee in industry. *J. Bus. U. Chicago*, 1955, 28, 58-66.
- OPINION RESEARCH CORPORATION. *The conflict between the scientific mind and the management mind*. Princeton: Opinion Research Corporation, 1959.
- PELZ, D. C., MELLINGER, G. D., & DAVIS, R. C. *Interpersonal factors in research, part 1: Studies in selected aspects of performance, communication and attitudes*. Ann Arbor: University of Michigan, 1954.

(Early publication received July 7, 1964)

AN INVESTIGATION OF THE CRITERION PROBLEM FOR A MEDICAL SCHOOL FACULTY¹

CALVIN W. TAYLOR, PHILIP B. PRICE, JAMES M. RICHARDS, JR.,

AND TONY L. JACOBSEN²

University of Utah

Through interviews, questionnaires, compendium listings, medical college files, and peer evaluations 80 criterion measures of on-the-job performance of 102 full-time college of medicine faculty members were developed and factor analyzed (a) to determine to what degree premedical and medical school grades predicted the criteria obtained, and (b) to explore other characteristics not currently emphasized in medical education. As indicated by the fact that 1 independent index of performance was obtained for approximately every 3 measures of performance included in the analysis, the most important finding was the complexity of the criterion area studied. School grades came out as a factor independent of the factors related to physician performance.

There can be little doubt that the most crucial person in any educational program is the teacher. This fact has been recognized in a large number of past studies of teacher performance. These studies have revealed some bias, however, in that they have involved almost exclusively elementary and secondary school teachers. A much smaller number of studies have been made of college teachers, and almost no studies have been made of the performance of teachers in graduate and professional schools. Yet, under the impact of problems increasingly facing our society, graduate and professional education is becoming more and more crucial. There is therefore a strong need for knowledge about the dimensions and correlates of the performance, accomplishments, and contributions of graduate and professional educators. The goal of this present study is to provide some of this needed knowledge for medical education.

The basic assumption of this study is that the performance of medical educators is complex and multivariable, and accordingly the procedure was based on that used by

Taylor and his associates (1958, 1961, 1963) in an investigation of the criterion problem for a group of physical scientists. In the study of physical scientists, more than 50 different measures of on-the-job performance were factor analyzed to isolate 17 different aspects of the productivity, creativity, and other contributions of the scientists. Similarly, in the present study a large number of measures were obtained of the on-the-job performance of the members of the full-time faculty (i.e., physicians who held clinical appointments and were primarily practitioners were excluded) of the University of Utah College of Medicine. These measures were then intercorrelated and factor analyzed to isolate several independent indices of performance.

METHOD

Criterion Measures. Eighty scores relevant to physician performance obtained from a variety of sources, including two scores measuring performance in education, were analyzed. The first 18 of these scores were obtained from the files, staff, and students of the college of medicine, including 5 scores obtained from official records, 9 scores obtained from the curriculum vita of each faculty member, 1 score based on listings in honorary compendiums, such as *Who's Who in American Medicine*, 1 score based on peer nominations as an outstanding contributor, 1 score based on student nominations as an outstanding teacher, and 1 score based on a rating by medical school department heads for "clinical excellence." The next 50 scores were obtained during an interview with each faculty member, and included 1 score based

¹ The research reported herein was supported through the Cooperative Research Program of the Office of Education, United States Department of Health, Education, and Welfare (Robert Beezer, Monitor).

² The authors would also like to express their appreciation to Louis D. Gusto, Theodore M. Yellen, Garry Shirts, and Gary Jorgensen for their assistance on this project.

on income, 2 scores having to do with society memberships, 37 scores based on answers to direct questions asked in the course of the interview, and 10 scores from a special questionnaire dealing with sources and degrees of occupational satisfaction. The final 12 scores were obtained from heterogeneous sources, and included 2 ratings by the project researcher who conducted the interview, 1 score involving ratings by expert judges of "overall performance" as indicated by all of the above 70 scores, 1 score indicating the degree to which each physician willingly participated in this research, 6 "control" scores involving such variables as years of experience, and 2 scores on performance in education, 1 as an undergraduate and 1 as a medical (or graduate) student.³

Subjects. The population studied consisted of the 118 members of the full-time faculty of the University of Utah College of Medicine during the 1961-62 academic year who had earned a doctoral degree (PhD, MD, or both). Letters signed by the Dean of the College of Medicine were sent to these 118 faculty members requesting them to participate in this project, and to grant an interview lasting approximately 1 hour. A second letter was sent to those faculty members who did not respond to the first letter, and a phone call was made until each faculty member either agreed or definitely refused to participate. Ultimately there were 102 faculty members who agreed to participate and this smaller group, therefore, was the sample actually studied in detail in this research. Within this group of 102, there were 73 participants with an MD degree and 29 participants with a PhD degree.

Scores were available for the remaining 16 non-interviewed faculty members on the 5 scores obtained from official records, the 6 control scores, the score based on citations in honorary compendiums, the rating by the medical school department heads, and the peer and student nomination scores. On these 15 scores, *t* tests were made comparing the group who did not participate in the interviews with the group who did participate. Only one significant difference was found, indicating that the group which did not participate obtained their first doctoral degree at a younger age than did the group which did participate. Since the single significant difference is on a control variable, it appears that the sample of 102 who did participate is not seriously biased.

Procedure. Scores on all variables, with the exception of the two dichotomous control variables dealing with type of faculty appointment and possession of an MD degree, were converted to normalized standard scores (Guilford, 1956, pp. 494-501) with a mean of 50 and a standard deviation of 10. The initial step in the data analysis was

to compute the 3,160 intercorrelations among the 80 scores.⁴ The resulting correlation matrix was then factor analyzed using the principal components solution based on eigenvalue and eigenvector analysis (Harmon, 1960). Unity was placed in the diagonal cells of the factor matrix, and all factors having an eigenvalue greater than 1.00 were extracted. These factors were then rotated to a final solution on the computer, using the varimax analytic rotational procedure.⁵

Since the computer factor analysis program used does not allow for missing scores, the mean score, or 50, was substituted for all missing scores. The effect of such a substitution is to reduce the correlation between variables and, therefore, in combination with unity in the diagonal, it produces some slight bias toward unique factors. In the present study, however, the only variables where there is much possibility of more than a minimal effect are those where the entire PhD group has missing scores. Examples of such variables are the department head's rating of "clinical excellence" and the "total number of hospitals employed in during residency." Since the number of such scores is not large, it would appear that the results of this study are not affected materially by substituting the mean for missing scores.

RESULTS AND DISCUSSION

An unexpectedly large number of factors, namely 25, had an eigenvalue greater than 1.00 and were included in the analysis.⁶ For these 25 factors the communalities ranged from .67 to .93 with an arithmetical average of .79. It was found that 54% of the communalities were of a magnitude of .80 or greater. In terms of factor loadings, the highest single loading per factor ranged from .50 to .92 with a mean loading of .79. The second highest loading per factor ranged from .32 to .90 with an average loading of .59. These 25 rotated factors are described below.

Factor A can be described best as Academic Seniority. In comparison with the other factors, it is a relatively general one, since more than half of the 80 variables have loadings greater than .20 on this factor. This factor

⁴ All computations for this project were carried out at the Western Data Processing Center, University of California, Los Angeles.

⁵ The rationale for this factor and rotation procedure is presented in detail by Kaiser (1960).

⁶ Copies of the complete correlation matrix, the unrotated factor matrix, and the rotated factor matrix are available by contacting Calvin W. Taylor, Department of Psychology, University of Utah, Salt Lake City, Utah.

³ The titles, sources, and details of the system by which raw scores were derived are available through the United States Office of Education (Price, Taylor, Richards, & Jacobsen, 1963) or by contacting Calvin W. Taylor, Department of Psychology, University of Utah, Salt Lake City, Utah.

includes those variables which relate to success and academic rank among the medical school faculty. Chief among these variables are age and experience. In general, one does not rise in academic rank without a certain degree of maturity and several years of various experiences. But mere age and experience are not the only correlates with academic rank. To achieve rank one must be producing scientifically, conducting research, publishing papers in scientific or professional journals, and attending and reading papers at scientific or professional meetings. Certain responsibilities will come with rank: a ranking member of the faculty will belong to and chair more committees in the medical school. But there will also be rewards and recognitions which go with academic rank, such as greater net income from the medical profession, than those of lesser rank. He will more likely be invited to serve as an editor for scientific journals or on scientific advisory boards. He will find it easier to obtain more research grants than those of lesser academic rank—or he who gets promotions is more capable of getting research grants. Students will tend to regard him as a better teacher. And he will also be seen by his colleagues as having made a significant contribution to medicine.

Factor B appears to involve Professional Recognition for Achievements. Persons scoring high on this variable have received a number of prestigious honors, and are nominated as outstanding by both peers and students. They have a higher than average output of written material, especially in recent years. Since their work has been written up in mass media more often than average, there is a slight indication that they have worked primarily in popular areas, or have done work in areas that have caught popular attention.

Factor C involves a greater than average use of scientific journals, but a smaller than average use of other techniques as a secondary method of keeping abreast. Persons scoring high tend to be less experienced, but more productive, and are held in higher esteem by their colleagues. Overall, the pattern of loadings suggests a “comer,” i.e., a new researcher and scholar who has not yet arrived but shows promise of doing so. While this is a

pattern which is difficult to summarize with a terse title, one possibility might be, Regular Review of Scientific Literature. This “well-readness” appears to involve basic sciences to a greater degree than more applied research (in such areas as clinical techniques and methodology).

Factor D is characterized mainly by its two largest loadings which indicate that persons scoring high on this factor received fellowships during their training and finished their training rapidly and therefore young. In other words, as students they effectively manipulated their academic situation, although they did not get particularly high grades. Considering all these results, this factor might most appropriately be titled Early Academic Progress and Recognition. However, there is some indication that high scorers have also been able to make later contributions, since they are awarded a greater than average number of relatively large research grants and are viewed as outstanding, both by their colleagues and by a panel of expert judges.

Factor E appears to involve Contributions as an Editor, since the most outstanding characteristics of high scorers are that they have been offered and have accepted a greater than average number of journal editorships. As might be expected, they derive a greater than average amount of satisfaction from writing and from reading. This suggests that they would also be high scorers on writing and reading communication abilities (Taylor et al., 1963). Contributions as an editor are regarded as important, since high scorers are both nominated by their peers as important contributors and rated as successful by expert judges. On this factor, there is, of course, a possibility of experimental dependence, since one must have been asked to be a journal editor before one can accept such an invitation. In view of the other high loadings on this factor, however, the authors do not feel that it could plausibly be described entirely or even largely as an artifact of including in the analysis both invitations to serve as an editor and acceptances of such invitations.

Factor F involves Self-Evaluated Accomplishment in Research since the variable

with the highest loading indicates that faculty members scoring high on this factor gave a larger than average number of research contributions when asked what their most important contributions have been. As with most self-evaluations, there appears to be considerable correspondence to reality, since these contributions are judged by others to be high quality, and high scorers get research grants and serve on advisory boards. As might be expected, high scorers on this factor tend more to have PhD degrees than MD degrees.

Factor G has its highest loading on the variable measuring the degree to which the faculty member himself estimates that his future publications will be comprehensive or full-scale reports rather than brief reports or abstracts. This is, no doubt, related to the fact that they have an above average number of research grants, and that these grants are relatively large ones. High scorers on this factor have tended to be the junior authors on past publications. They are, however, considered to be outstanding contributors by their colleagues. This factor might best be summarized as Self-Estimated Comprehensiveness of Future Publications.

Factor H is concerned with Research Consulting, and somewhat more informal consulting than formal. This suggests that the high scorer is approachable, and quite willing to talk to and advise other people about research problems. This is true in spite of the fact that the high scorer himself has been more productive in research and presumably busier. The high scorer is more likely to be a PhD and is consulting as a basic scientist.

Factor I is characterized by a high negative loading on liking for actual practice as reported on a self-rating. The best title for this factor might be Rejection of Actual Practice. Persons scoring high on this factor tend to dislike teaching, have a large number of relatively big research grants, and expect to produce many publications in the immediate future although they have not given any particular evidence of productivity in the past. There is some suggestion of a person who is retreating to the laboratory because of a dislike of activities requiring a great deal of interpersonal contact.

Factor J might best be doubly titled Prolonged Residency Training—Excellence as a Clinician. High scorers have a more varied and longer than average background of hospital experience, and are considered to be outstanding clinicians by their department heads. They are frequently called into medical consultation, but are less productive of publications. Thus, their orientation is primarily that of a physician rather than a scientist or teacher, even though they are full-time faculty members. It is obvious that a college of medicine needs some persons with this orientation among its staff.

Factor K can best be described as Level of Medical Specialization and Attainment. Persons scoring high tend to have a more than average number of high quality specialty certifications. They have spent more (prerequisite) years in internship and residency than average. The fact that they have a larger than average number of listings in honorary compendiums is at least partly due to the fact that passing a specialty board is a requirement for being included in some of the compendiums used. Persons scoring high on this factor also tend to derive more satisfaction from teaching than from research. There is some evidence in the lack of sizable loadings for variables having to do with research and teaching productivity that some high scorers used up much of their psychic energy getting their speciality certifications and consequently are just keeping things going now in their highly specialized area.

Factor L has its highest loadings on variables having to do with clinical teaching and consulting on medical problems. As is to be expected, persons scoring high on this factor tend to have a MD degree. Since there are no large loadings on variables having to do with research, the overall picture suggested by this factor is that it describes a person who focuses on transmitting known answers rather than searching for new (or better) answers. Thus, persons scoring high on this factor are serving a function vital to all of education (Taylor et al., 1962) of "bridging the gap" or communication between basic research and practice. This type of activity is liked by medical students, and is rewarded with a higher than average income. The high load-

ings for consulting activities suggest that this factor also involves a form of visibility within the medical faculty, or perhaps a good professional appearance. Oral expression is a common thread running through most variables loading on this factor. Overall, the most obvious title for this factor would be Clinical Teaching and Consulting.

Factor M appears to be Academic Orientation—Teaching Excellence. High scorers have academic rather than research appointments, are responsible for a large number of courses, and are considered to be outstanding teachers by their students. They are not outstandingly productive of publications and are regarded as average clinicians. Thus, they are making their contributions primarily as teachers rather than as scientists or physicians. While this type of contribution is not particularly considered to be unimportant by colleagues in the college of medicine, neither is it considered to be particularly outstanding.

Factor N pretty clearly involves Teaching Responsibilities, since the variables with the highest loadings are those dealing with course responsibilities. High scorers also have more positions on advisory boards. They are regarded as better than average clinicians by their department heads, but are viewed as neither outstandingly good nor outstandingly bad by their students. Overall, the high scorer seems to be a person who is fully occupied in teaching and therefore makes only average contributions in, for example, research.

Factor O appears to involve Administrative Contributions to the College of Medicine, since its highest loadings indicate that high scorers derive a greater than average amount of satisfaction from miscellaneous chores (which are necessary but not particularly popular or prestigious) and have been more active as committee chairmen. The contributions of the high scorer are primarily local, since he has a smaller than average number of memberships and offices in professional societies. It appears, therefore, that this factor is somewhat similar to the Current Organizational Status factor obtained in the first author's study of physical scientists (Taylor, et al., 1961).

High loadings on Factor P indicate rapid promotion in the medical school, and a lot of moving from position to position without staying very long in any one position. Since no single thing (such as research or teaching) distinguishes the high scorer on this factor, there is some suggestion of a man who is "playing the ropes," and who is willing to do whatever the organization requires to get ahead. Therefore, an appropriate title for this factor is Rate of Professional Mobility and Advancement.

Factor Q is complex and somewhat difficult to interpret. The most salient points appear to be that high scorers have held a greater than average number of society offices and committee chairmanships, and tend to respond to a question about their most important contributions with things which are not closely related to medicine. They are not particularly productive in research, and are regarded as poorer than average teachers by students. Thus, the only thing that distinguishes them is that they have somehow attained positions with relatively high status, while having done little to justify this status. This suggests that the best title for this factor is Status Seeking Tendencies and Skills.

Factor R might best be interpreted as Public Recognition for Contributions since its highest loadings are on number of speeches to laymen groups and number of times research has been written up in mass media. The high scorer also belongs to a large number of social organizations. This pattern suggests that the high scorer enjoys moving among nonmedical people, and is representing the medical profession to a lay audience. He has produced a relatively large number of papers in the last few years, and has tended to be sole or senior author, but derives a smaller than average amount of satisfaction from research. This raises the possibility that he is using his research to seek status more than knowledge, and status more outside than within his profession.

Factor S appears to involve Participation in Professional Societies. Those scoring high like to participate in societies and go to a lot of society meetings. However, the societies they belong to are relatively low in prestige.

They also express a liking for consulting, but are regarded as below average clinicians by their department heads. The overall pattern suggests that high scorers are persons who are striving to maintain their own visibility within the profession.

Factor T has high loadings on memberships and offices in civic and political organizations, and, accordingly, the most obvious title is Civic Participation. Although one must be a member of an organization before one can be an officer, there is little reason to believe in the present case that inclusion of both variables involved experimental dependence to a degree which would produce an artificial factor. Since in absolute numbers no one person on the faculty held offices in a large number of civic organizations, such experimental dependence would have a serious effect only if there were a larger proportion of the faculty with no memberships in civic organizations and therefore no offices. Such is not the case. This is, of course, a side activity for faculty members, and in this connection it is interesting to note that high scorers on this factor tend to have produced a lower than average number of recent publications and are considered to be below average clinicians. Thus there is a suggestion of a man who has achieved little status in his college of medicine, and as a result is seeking civic and political status outside of his college.

Factor U should be titled Overall Occupational Satisfaction, since the variables with the highest loadings are Overall Occupational Satisfaction and three self-ratings of Liking for Research, Administration, and Teaching. The opposite end of the factor is negative and gives the overall picture of an unhappy man who has not found a suitable niche for himself in life. The most striking thing about this factor is that the Overall Occupational Satisfaction score is factorially simple, appearing only on this factor, and therefore quite independent of other aspects of performance. If the finding of this study should prove to be generally true, that satisfaction with a vocation is independent of performance in that vocation, far reaching implications for vocational counseling would follow.

Factor V involves Participation in Social

Organizations with its two highest loadings on the variables having to do with participation in social organizations. As was the case with the Civic Participation factor, the authors do not believe that emergence of this factor can be attributed solely to the experimental dependence of holding office in a social organization on being a member of that organization. Persons scoring high also tend to have a large number of society offices, but in societies with relatively low status. This, in combination with negative loadings for productivity of publications, suggests that the high scorer may be somewhat of a "glad-hander," spending a considerable amount of his energies in social life so that he has less energy to expend in being a scholar or researcher.

Factor W mainly involves the impact the faculty member had on the person who interviewed him for this research. Specifically, the project interviewers generally rated individuals characterized by this factor as quite likeable. It would seem to follow that these physicians move easily among people and have considerable social skills. Thus, an appropriate label for this factor is that of Interviewer's Rating of Likeability. Also, it is interesting to note that this factor is independent of Factor V, Off-the-Job Socializing. It is commonly assumed that persons who are socially adept in one situation will be so in others. However, the results obtained with this sample suggest that what has been taken to be the general ability of gregariousness or likeability is actually more than one attribute, each of which may be somewhat situation-bound.

Factor X has its highest loading on the variable measuring how easy it was to schedule the research interview. An obvious interpretation, therefore, is as Project Cooperativeness. Secondary loadings indicate that the high scorer has had his work written up in mass media, has considerable course responsibility, is frequently cited in honorary compendiums, and has a greater than average number of research grants. This pattern suggests a person who has arrived, and is therefore secure enough to cooperate willingly with a project in which his performance is evaluated.

Factor Y clearly concerns Achievement in Education: its two highest loadings are on undergraduate school grades and grades achieved in medical (or graduate) school. The secondary loadings show that persons scoring high tend to dislike consulting and administrative activities and whatever research grants they now have are relatively large. However, and most significantly, high scorers have not distinguished themselves in anything else; the 24 other factors obtained in this study, including those measuring the more important professional contributions and accomplishments, are essentially unrelated to this dimension of academic achievement. In this regard, it should be emphasized that the finding of an educational performance factor independent of professional performance is not to be interpreted as an artifact of the use of an orthogonal rotation. In other words, emergence of an independent performance in education factor seems to be a real finding, and one of considerable and obvious importance.⁷ It should also be emphasized that it would be difficult to justify the traditional liberal arts argument that education is a worthy activity in and of itself whether or not it is related to other attainments for professional education such as medical school. These results, therefore, suggest the criteria currently used to evaluate medical students should be searchingly examined.

Taking this study as a whole, perhaps the single most important finding was the great complexity of the total criterion area for the college of medicine faculty (this was certainly more important than the specific details of the factors obtained). There was not a large amount of overlap, generally, between the different measures of physician performance. An indication of this complexity is given by the fact that one factor was ob-

tained for approximately every three measures of performance included in the analysis. This means that one should not attempt to measure performance by one score such as number of publications, or even by a small number of scores (or, for that matter, a single factor or small number of factors), since then the criterion problem may appear much less complex than it really is. As soon as a large number of measures of performance are obtained, the real complexity of the criterion problem emerges very forcefully. However, within this complexity the rotated factor solution provides a powerful and interesting analytical view of the criterion problem as well as a more clear and sound basis for organizing the measures of physician performance.

It appears that there are many outlets in which a faculty member can expend his energies and efforts. Thus, an individual faculty member can focus his efforts into a larger or smaller number of these efforts, but it is unlikely that an individual can devote an above average amount of energy to all of the available outlets. As a result, both the selection of outlets from the total number available and the effectiveness with which one's efforts are put to use in these outlets are important in assessing the overall accomplishments and contributions of an individual. Both of these considerations should be used in evaluating current faculty members for retention and promotion and in evaluating the experience and attainment elsewhere of candidates for appointment to the faculty.

The criterion complexity revealed by these results has definite implications for the prediction of the performance of medical faculty members. It is unlikely that many of these criteria will be highly correlated with the usual predictors, such as grades and typical aptitude tests, which almost exclusively involve verbal ability. This is confirmed by the fact that, in this study, performance as a student in medical school came out as a factor independent of all of the factors related to performance as a medical faculty member. On the other hand, recent research (Ellison & Taylor, 1962) indicates that biographical inventories can be used suc-

⁷ We do not make this point without being aware of various arguments that some might raise in objection to it, such as, that the lack of correlation shown between academic performance and the on-the-job performance measures is due to "restriction of range." A lengthy, 11-point technical discussion explaining why the authors feel such arguments, at best, can only account for a small portion of this lack of correlation is presented elsewhere (Price, Taylor, Richardson, & Jacobsen, 1963).

cessfully in the prediction of multiple independent criteria, yielding cross-validities ranging from .40 to .60 against a variety of measures of the creativity and productivity of scientists. In the opinion of the authors, self-assessments are among the types of predictor most likely to yield useful validities against the criteria isolated in this study.

REFERENCES

- ELLISON, R. L., & TAYLOR, C. W. The development and cross-validation of a biographical inventory for predicting success in science. *Amer. Psychologist*, 1962, 17, 391. (Abstract)
- GUILFORD, J. P. *Fundamental statistics in psychology and education*. New York: McGraw-Hill, 1956.
- HARMON, H. H. *Modern factor analysis*. Chicago: Univer. Chicago Press, 1960.
- KAISER, H. F. The application of electronic computers to factor analysis. *Educ. psychol. Measmt.*, 1960, 20, 141-151.
- PRICE, P. B., TAYLOR, C. W., RICHARDS, J. M., JR., & JACOBSEN, T. L. Performance measures of physicians. (Contract No. OE-2-10-093) Washington, D. C.: United States Department of Health, Education, & Welfare, 1963.
- TAYLOR, C. W., GHISELIN, B., & WOLFER, J. Bridging the gap between basic research and educational practice. *NEAJ.*, 1962, 51, 23-25.
- TAYLOR, C. W., SMITH, W. R., & GHISELIN, B. The creativity and other contributions of one sample of research scientists. In C. W. Taylor and F. Barron (Eds.), *Scientific creativity: Its recognition and development*. New York: Wiley, 1963. Pp. 53-76.
- TAYLOR, C. W., SMITH, W. R., GHISELIN, B., & ELLISON, R. L. Explorations in the measurement and prediction of contributions of one sample of scientists, *USAF ASD tech. Rep.*, 1961, No. 61-96.
- TAYLOR, C. W., SMITH, W. R., GHISELIN, B., SHEETS, B. V., & COCHRAN, J. R. Identification of communication abilities in military situations. *USAF WADC tech. Rep.* 1958, No. 58-92.

(Received June 20, 1963)

ESTIMATING THE NUMBER OF DIFFERENT SELECTION DECISIONS RESULTING FROM THE USE OF ALTERNATE PREDICTOR COMPOSITES

WARREN W. WILLINGHAM

Georgia Institute of Technology

A method was described whereby d , the number of different selection decisions resulting from the use of alternate predictor composites may be rapidly estimated from the proportion of applicants accepted and the correlation between the composites. It was further demonstrated that the correlation between 2 optimally weighted composites based upon m and n variables ($m \subset n$) is equal to the ratio of the 2 multiple correlations. 2 examples were used to illustrate the use of d in evaluating selection strategies.

There have been notable advances in the technology of evaluating the cost and efficiency of alternate selection strategies (Brogden, 1949; Cronbach & Gleser, 1957; Taylor & Russell, 1939). Effective communication of such evaluations to managerial and administrative personnel is necessarily an immediate concern of the investigator. Psychologists have resorted to a variety of bar graphs, expectancy charts, et cetera, in an effort to paint a comprehensible picture of a technical problem.

One piece of information which is seldom reported is the actual number of selection decisions which would be changed if one predictor composite were used in lieu of another. Carrying it a step further, what sort of criterion performance would we expect from those in-

dividuals accepted on one composite but not on another? This type of information is of particular interest to a person who has administrative responsibility for selecting applicants since he frequently perceives his problem largely in terms of marginal applicants. One possible reason why investigators seldom use this method of explaining the effect of using alternate batteries is the fact that the necessary information is tedious to obtain. The purpose of this article is to (a) describe a rapid method for estimating the number and relative criterion performance of applicants selected by one battery but not the other and (b) demonstrate a simple relationship between number of altered selection decisions and increments in the multiple correlation.

METHOD

The direct method of determining the number of different selection decisions with a short and a longer battery would involve computing a predicted criterion score based upon m variables, computing another predicted score based upon n variables ($m \subset n$), arranging the applicants in rank order on the two scores, designating a cutting score, and actually counting the number of discrepant decisions. Figure 1 illustrates the basis of a much simpler analytic method for estimating the number of such discrepancies. The figure shows a schematic scatterplot of the original predictor composite (C_I) against the composite (C_{II}) augmented by additional predictors. A cutting score has been set arbitrarily at the thirtieth percentile of each composite. The same selection decision would be reached for all applicants in the regions marked S, regardless of which battery were used. The regions D_1 and D_2 include those individuals for whom a different decision would be made. Assuming the joint distribution of C_I and C_{II} is normal bivariate, $D_1 = D_2$ and the proportion of applicants in these areas can be determined from exist-

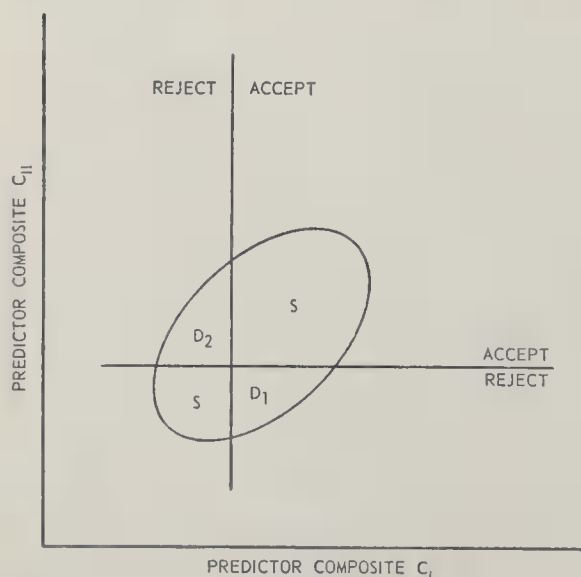


FIG. 1. Schematic diagram of accept-reject regions using alternate predictor composites.

ing tables (Lee, 1927). All that is necessary is P_a , the proportion of applicants accepted, and r_c , the correlation between C_I and C_{II} . The former is given; a simple method for determining r_c is described in the following section.

Tables of the normal correlation surface are quite extensive and not readily obtainable. Table 1 includes the necessary entries expressed in a more convenient form for the purpose at hand. Table 1 is entered with r_c and P_a in order to determine d , the proportion of all applicants who fall in either of the regions D_1 or D_2 . Consequently

dN = the number of applicants accepted by C_{II} who would have been rejected by C_I ,

$2dN$ = the total number of altered decisions, and

d/P_a = the proportion of accepted applicants affected by the choice of C_I or C_{II} ,

where N equals the total number of applicants. Furthermore, it is possible to estimate the differential probability of success of those individuals who fall in areas D_1 or D_2 . The Taylor-Russell tables (1939) can be used to estimate the overall proportion of successful applicants that would be accepted using either C_I or C_{II} . The difference in those proportions of successful applicants $[P_{s(II)} - P_{s(I)}]$ must be accounted for by the dN individuals accepted by C_{II} but rejected by C_I . Therefore, the difference in the probability of success of those applicants in areas D_1 and D_2 is equal to $[P_{s(II)} - P_{s(I)}]P_a/d$.

Determining r_c

The method described above is generally applicable to any two predictor composites. It is actually irrelevant whether the variables are optimally weighted, whether one or more variables are included in each composite, or whether the m variables of one composite are all included among the n variables of the other. The correlation between two linear composites can be determined from summary statistics which are normally available, but the most efficient method of computing r_c will depend upon the nature of the composites.

Statistics and measurement texts frequently report equations for part-whole correlation and the correlation between composites, but in the general case r_c can be determined most easily by manipulating the variance-covariance matrix. This extremely flexible and simple method of computing correlations between weighted composites has been discussed in detail by DuBois (1957). Perhaps the most common situation in which it may be desirable to compute r_c is when an investigator wishes to evaluate the effect of adding one or more variables to a regression equation. In this case, the equation for r_c reduces to a very convenient expression.

Given a criterion x_0 , a predicted criterion score C_I based upon m optimally weighted predictors, and a predicted criterion score C_{II} based upon n optimally weighted predictors ($m \subset n$), the correlation between C_I and C_{II}

$$r_c = \frac{\text{Cov}_{I \text{ II}}}{\sigma_I \sigma_{II}}.$$

TABLE 1

THE PROPORTION D OF ALL APPLICANTS WHO ARE
ACCEPTED ON ONE COMPOSITE BUT
REJECTED ON THE OTHER

r_c	Percentage of applicants accepted					
	5- 95%	10- 90%	20- 80%	30- 70%	40- 60%	50%
.00	.047	.090	.160	.210	.239	.250
.05	.047	.088	.156	.203	.232	.242
.10	.046	.087	.152	.197	.224	.234
.15	.045	.085	.147	.191	.217	.226
.20	.045	.083	.143	.185	.209	.218
.25	.044	.081	.138	.178	.201	.210
.30	.043	.078	.134	.171	.193	.202
.35	.042	.076	.129	.165	.186	.193
.40	.041	.073	.123	.158	.178	.185
.45	.039	.070	.118	.150	.169	.176
.50	.038	.067	.113	.143	.160	.167
.55	.036	.064	.107	.135	.151	.157
.60	.034	.061	.100	.127	.142	.147
.65	.032	.057	.094	.118	.132	.137
.70	.030	.053	.087	.109	.122	.127
.75	.028	.049	.079	.099	.111	.115
.80	.025	.044	.071	.088	.099	.102
.85	.022	.038	.061	.076	.085	.088
.90	.018	.031	.050	.062	.069	.072
.95	.013	.022	.035	.044	.049	.051
1.00	.000	.000	.000	.000	.000	.000

Note.—Entries refer to either, but only one, of the areas D_1 or D_2 of Figure 1.

Without loss of generality all primary variables can be expressed as z scores, in which case

$$\sigma_I = R_{0(12...m)} \quad \text{and}$$

$$\sigma_{II} = R_{0(12...n)}$$

The covariance may be expressed in matrix notation as

$$\text{Cov}_{I \text{ II}} = V_m' E_{mn} D_n 1$$

where V_m' is a row vector of standard regression weights based upon m predictors, D_n is a diagonal matrix of standard regression weights based upon n predictors, and E_{mn} is a horizontal matrix consisting of the first m rows of the n order correlation matrix. The product $E_{mn} D_n$ may be regarded as a column supervector whose successive subvectors have elements which are seen to be identical to the first n terms of the first m simultaneous equations defining D_n . Consequently, the product $E_{mn} D_n 1$ is equivalent to V_v , an m order column vector of validity coefficients. Since it is well known that the minor product

$$V_m' V_v = R_{0(12...m)}$$

we can write

$$r_c = \frac{R_{0(12...m)}}{R_{0(12...n)}}.$$

When n includes m , the correlation between the predicted score based upon m variables and the predicted score based upon n variables is equal to the ratio of the respective multiple correlations.¹

EXAMPLE 1

Company X is instituting a program for selecting salesmen. Crossvalidated results on unselected new salesmen indicated that a short intelligence test plus number of years of selling experience would yield a multiple correlation of .45 with the criterion. A biographical blank requiring 1 hour to administer and score would raise the multiple to .50. The company expects to hire 140 salesmen a year from among approximately 200 applicants. In the unselected group, 30% of the new salesmen were judged to be unsatisfactory. The immediate concern is whether or not to include the biographical blank.

Entering Table 1 with $r_c = .90$ and $P_a = .70$, we find $d = .062$. Thus, including the biographical blank would mean that approximately $dN = 12$ different applicants would be accepted, or 9% of all accepted applicants. From the data given, the Taylor-Russell tables (1939) indicate that 78% of applicants accepted on the two variable composite would be satisfactory salesmen, as opposed to 80% of those selected on the three variable composite. Consequently there would be an expected difference of 22% (.02/.09) in the success rate of those individuals differentially selected by the alternate batteries. Expressed in absolute terms, using the longer battery

¹ After this manuscript was accepted for publication, the writer found a more general proof for this relationship which does not assume that the m variables of C_I are optimally weighted. So long as the n variables of C_{II} are optimally weighted and n equals or includes m , r_c can be expressed as the ratio of the validities of C_I and C_{II} . This is a fortunate result since it facilitates the comparison of old versus new regression equations or optimal versus approximate weights.

would result in accepting 12 applicants in lieu of 12 other applicants, with a net gain of approximately three successful salesmen. Whether the gain warrants the extra effort would depend, of course, upon many additional conditions.

EXAMPLE 2

Administrators are sometimes prone to select applicants on the basis of the unweighted sum of two superficially valid predictors, one of which adds little or nothing to the other. The notion of altered decisions can be useful in illustrating the futility of this procedure.

Assume two predictors intercorrelate .60, have validities of .30 and .50, and for the sake of convenience, $\sigma_1 = \sigma_2$. In this case the formula for part-whole correlation reduces to $[(1 + r_{12})/2]^{\frac{1}{2}}$ so r_c , the correlation between (x_1) and $(x_1 + x_2)$, equals .90. With a selection ratio of .50, about one out of seven selection decisions would be changed by using the composite instead of the better predictor—but with no gain in the proportion of successful acceptees. If the two predictors were optimally weighted, decisions based upon the better predictor or upon both predictors would be identical since the less valid predictor would have a weight of zero.

REFERENCES

- BROGDEN, H. When testing pays off. *Personnel Psychol.*, 1949, 2, 171-185.
- CRONBACH, L. J., & GLESER, GOLDINE. *Psychological tests and personnel decisions*. Urbana: Univer. Illinois Press, 1957.
- DuBOIS, P. *Multivariate correlational analysis*. New York: Harper, 1957.
- LEE, ALICE. Supplementary tables for determining correlation from tetrachoric groupings. *Biometrika*, 1927, 19, 354-404.
- TAYLOR, H. C., & RUSSELL T. J. The relationship of validity coefficients to the practical effectiveness of tests in selection. *J. appl. Psychol.*, 1939, 23, 565-578.

(Received July 16, 1963)

JOB ATTITUDES IN MANAGEMENT:

VI. PERCEPTIONS OF THE IMPORTANCE OF CERTAIN PERSONALITY TRAITS AS A FUNCTION OF LINE VERSUS STAFF TYPE OF JOB¹

LYMAN W. PORTER AND MILDRED M. HENRY

University of California, Berkeley

In a questionnaire study, over 1800 managerial respondents rank-ordered 5 Other-Directed or Organization Man personality traits and 5 Inner-Directed traits in terms of their importance for job success. Responses were tabulated by 3 types of managerial positions: line, combined line-staff, and staff. Results showed that staff managers placed relatively more emphasis on the Other-Directed traits and less emphasis on the Inner-Directed traits than did line managers. Managers in combined line-staff jobs were intermediate between the other 2 groups in their responses.

A previous paper reported on managerial perceptions of the importance of certain personality traits for "job success" as a function of level of position within management (Porter & Henry, 1964). The present paper deals with similar perceptions as a function of line versus staff types of managerial positions. Specifically, this study investigates line-staff differences in the relative importance of Inner-Directed type personality traits versus Other-Directed traits for job success. There has been little previous empirical research directed to this problem, although some writers in the popular press, notably William H. Whyte Jr. (1956), have made rather definitive assertions in this area. Whyte, for example, in his book, *The Organization Man*, implies in a number of places that the Other-Directed type of Organization Man aspires more often to staff jobs—especially jobs in such areas as personnel—than he does to line jobs. The present study attempts to ascertain whether managers who are actually working in line and in staff jobs feel that Other-Directed traits are more

necessary for success in staff jobs than they are for success in line jobs.

METHOD

Questionnaire

One section of a three-section questionnaire was used to collect the data for this study. The relevant section of the questionnaire contained the following instructions (in part):

The purpose of [this part of the questionnaire] is to obtain a picture of the traits you believe are necessary for success in YOUR PRESENT MANAGEMENT POSITION. Below is a list of 12 traits arranged randomly. Rank these 12 traits from 1 to 12 in the order of their importance for success in your present management position.

Although 12 traits were presented to the respondents for ranking, 2 of these traits were camouflage items that were included in the list to disguise the dimension being studied. These camouflage traits were Intelligent and Efficient, and the ranks given them by a respondent were not counted, and in the data tabulations they were removed from the list of ranks to prevent them from having any direct effect on the rankings of the 10 relevant traits. Thus, for each respondent the 10 relevant items were reranked after the 2 camouflage items had been removed. (Whenever a camouflage trait was removed, the ranks of the relevant items below it were all moved up one rank.)

The 10 relevant traits are listed in Table 1 in the Results section. Five of the traits are labeled Inner-Directed to designate one end of the composite dimension being measured, and the other 5 traits are labeled Other-Directed to designate the other end of the composite dimension.

As was explained in the previous article (Porter & Henry, 1964) using these 10 traits:

¹ This study was carried out as part of the research program of the Institute of Industrial Relations, University of California, Berkeley. The data were collected while the senior author was a Ford Foundation Faculty Research Fellow. The Institute of Social Sciences at the University of California and the American Management Association contributed to the support of the research assistance, and the Computer Center of the University provided facilities for data computation.

The personality trait dimension to be investigated here is intended to be a composite dimension made up of related but slightly different personality continua. One of these specific continua is Riesman's Inner-Directed—Other-Directed distinction (Riesman, 1950). Another is Whyte's description in *The Organization Man* of the Protestant Ethic versus the Social Ethic as behavior guides in management (Whyte, 1956). A third source of the composite dimension was the contrasting picture of self-descriptions of top managers versus middle managers in a previous article on self-perceptions (Porter & Ghiselli, 1957). Still another related continuum is one which has been termed the "entrepreneurial-bureaucratic" dimension. Considered together, then, these various specific continua listed above have much in common, but each is slightly divergent in emphasis from the others. The two clusters of traits used in the present study constitute an attempt to include aspects of each continuum in a composite dimension [p. 32].

The labels Inner-Directed and Other-Directed were attached to the two clusters of traits simply as a shorthand way of describing the dimension.

Procedure and Sample

To obtain the sample of respondents for this study the questionnaire was mailed to nearly 6,000 managers working in a wide variety of types and sizes of companies throughout the country.² Thirty-three percent of the managers responded to the questionnaire, and specifically, 1,896 filled out the particular section of the questionnaire used in this study; however, only the responses from the 1,786 managers below the President level were used for this study, since company presidents are not classifiable into specific line or staff positions. About two-thirds of the respondents were from manufacturing companies, and the remainder were from non-manufacturing firms.

The major independent variable investigated in this study was the horizontal location of positions—i.e., line versus staff distinctions. Respondents were placed into one of three categories of positions—line, combined line-staff, and staff positions—on the basis of their own self-classification. It should be emphasized that respondents carried out this self-classification into line or staff positions at the end of the questionnaire, where a number of personal data questions were asked concerning age, education background, type of company, department within company, etc. Thus, it is quite unlikely that in filling out the prior parts of the questionnaire, which included the ranking of the traits studied in this article, respondents were systematically distorting their answers based on whatever stereo-

types they might have held concerning line or staff positions. At the time they did their rankings they had no knowledge of the investigator's specific interest in the line-staff variable, or, for that matter, in any other particular independent variable.

Although the line-staff variable was the major independent variable studied, respondents were also cross-classified on another independent variable, level of position within management. This was done to control for any possible effect of level of position on the results for line versus staff comparisons, since previous studies (Porter, 1962; Porter & Henry, 1964) have demonstrated that the level variable has a strong relationship to job perceptions. Another advantage of cross-classifying by level was that four independent subsamples of respondents could be created to test for line-staff differences. The method of classifying respondents by management level has been described in detail in a previous paper (Porter, 1962). The four categories of this dimension used in the present study were: Vice-President, Upper-Middle, Lower-Middle, and Lower.

Information on age and amount of formal education was also obtained for each of the three major groups of respondents in the present study—line, combined line-staff, and staff managers. Tabulations on these two variables showed highly similar median values among the three groups.

RESULTS

The basic results of the study are presented in Tables 1 and 2. Both tables deal with each of the 10 traits separately and also with the division of the traits into Inner-Directed and Other-Directed clusters of 5 traits each.

Table 1 is designed to show general trends of differences among the three types of positions—line, combined line-staff, and staff—in the mean importance attached to each of the 10 traits and the two clusters of traits. The values for the individual traits in Table 1 were obtained by taking the average of the mean ranks of each trait across three levels of management—Vice President, Upper-Middle, and Lower-Middle—for each of the three types of positions. (Mean ranks from the Lower level were not used since the combined line-staff category did not have a sufficiently large *N*.) That is, mean ranks for each trait were obtained for each of these three levels of management for each type of position. Then, the three mean ranks (obtained from the three management levels) were averaged for a given type of position, e.g., line positions, to obtain the values to

²The assistance of the American Management Association, and particularly Robert F. Steadman, in obtaining the sample of respondents is gratefully acknowledged.

TABLE 1

MEAN IMPORTANCE ($10 - \bar{X}$ rank) OF TRAITS FOR JOB SUCCESS BY TYPE OF POSITION

Trait	Type of position		
	Line ($N = 507$)	Combined Line- Staff ($N = 492$)	Staff ($N = 684$)
Inner-Directed			
Forceful	4.55	3.90	3.70
Imaginative	6.62	6.67	6.76
Independent	2.38	2.50	2.61
Self-Confident	5.70	5.57	5.43
Decisive	6.49	5.84	5.49
Total for cluster	25.74	24.48	23.99
Other-Directed			
Cooperative	5.45	5.77	5.73
Adaptable	4.98	5.05	5.03
Cautious	1.06	1.42	1.45
Agreeable	2.53	2.79	2.95
Tactful	5.26	5.51	5.85
Total for cluster	19.28	20.54	21.01

Note.—Higher numbers indicate greater importance and values are averages for Vice-President, Upper-Middle and Lower-Middle levels of management.

be used in Table 1. Before these values were placed in Table 1, however, they were subtracted from the constant number 10 so that

higher numbers in Table 1 represent greater perceived importance.

In effect, Table 1 allows for comparisons among the three types of positions with the variable of management level controlled. It can be seen in Table 1 that the importance of the Inner-Directed cluster of traits decreased from line, to combined line-staff, to staff jobs, and therefore the importance of the Other-Directed cluster showed a corresponding line-to-staff increase. It is also clear from Table 1, though, that not all traits contributed equally to these trends. In particular, two of the Inner-Directed traits, Imaginative and Independent, showed reverse trends. Among the five Other-Directed traits, Adaptable showed no appreciable trend in either direction.

Table 2 is designed to show with more precision the trends evident in Table 1 and especially to indicate the level of significance of the trends. The values in Table 2 were obtained by subdividing the sample by management level as well as by type of position. For this table, mean ranks were obtained for the three types of positions within each

TABLE 2

NUMBER OF CHANGES IN SIZE OF MEAN IMPORTANCE OF TRAITS FOR JOB SUCCESS: FROM LINE TO COMBINED LINE-STAFF TO STAFF TYPE OF JOBS WITHIN FOUR MANAGEMENT LEVELS

Trait	Management levels												Total sample		
	Vice President			Upper- Middle			Lower- Middle			Lower					
	+	0	—	+	0	—	+	0	—	+	0	—			
Inner-Directed															
Forceful	0	0	2	0	0	2	0	0	2	0	0	1	0	0	7*
Imaginative	1	0	1	1	0	1	2	0	0	1	0	0	5	0	2
Independent	1	0	1	2	0	0	1	0	1	1	0	0	5	0	2
Self-Confident	0	1	1	0	1	1	0	1	1	1	0	0	1	3	3
Decisive	0	0	2	0	0	2	0	0	2	0	0	1	0	0	7*
Total	2	1	7	3	1	6	3	1	6	3	0	2	11	3	21
Other-Directed															
Cooperative	1	0	1	1	1	0	1	0	1	1	0	0	4	1	2
Adaptable	1	1	0	0	1	1	1	0	1	0	1	0	2	3	2
Cautious	2	0	0	2	0	0	1	0	1	1	0	0	6	0	1
Agreeable	1	1	0	1	1	0	1	1	0	0	1	0	3	4	0
Tactful	2	0	0	2	0	0	2	0	0	1	0	0	7	0	0*
Total	7	2	1	6	3	1	6	1	3	3	2	0	22	8	5**

* $p < .05$.** $p < .01$.

of the four levels of management—Vice President, Upper-Middle, Lower-Middle, and Lower.³ (For one of the 12 subgroups thus obtained, combined line-staff managers at the Lower level, the *N* was less than 20, and therefore this subgroup was not included in tabulations based on the means of these subgroups. Other *N*s ranged in size from 41 to 291.) Table 2 is constructed to show the changes in the mean importance of each trait from line to staff jobs within each of the four management levels. Whenever a mean increased by more than .05 of a mean rank from a line to a line-staff position or from a line-staff to a staff position an increase or + was counted; when the mean decreased more than .05 of a mean rank a decrease or - was counted. Changes of .05 or smaller were recorded as no changes or 0. To illustrate, for the trait Cooperative: At the Vice President level the mean importance increased from the line to the combined line-staff group and decreased from the combined line-staff group to the staff group. Therefore, one + and one - were recorded under the Vice President column for Cooperative. At the Upper-Middle level, the importance of Cooperative increased once and showed no change once, in going from line to staff positions. At the Lower-Middle level there was one increase and one decrease, and at the Lower level (where there was no intermediate combined line-staff group because of an *N* less than 20) there was one increase. Adding across levels for Cooperative, as shown in the right-hand columns of Table 2, gives a total of four increases, one no change, and two decreases. This indicates an insignificant trend for Cooperative to be more important in Staff than in Line jobs.

The trends in Table 2, as summarized in the right-hand columns, were evaluated by the sign test. Table 2 shows that two Inner-Directed Traits, Forceful and Decisive, and

one Other-Directed trait, Tactful, produced significant trends in the expected direction of Line managers placing more importance on Inner-Directed behavior and less on Other-Directed behavior than Staff managers. Four other traits showed insignificant trends in the same directions, while Adaptable showed no trend and Imaginative and Independent showed nonsignificant reverse trends. Looking at the totals for the two clusters in Table 2 it can be seen that because of the results for Imaginative and Independent the overall trend existent in the Inner-Directed cluster failed to reach statistical significance. However, for the Other-Directed cluster there was a trend significant at the .01 level of confidence for staff managers to place more importance on these five traits considered as a total group.

It can also be seen from looking at the bottom row for each cluster in Table 2 that the overall trends for the Inner- and Other-Directed clusters held up *within each* of the four management levels, although there was a slight reversal in the Inner-Directed cluster at the Lower level.

One other aspect of the findings should be noted: As is directly observable in Table 1 and indirectly indicated in Table 2, managers in combined line-staff jobs are intermediate between managers in pure line jobs and those in pure staff jobs in the importance they attach to each cluster of traits. In other words, managers in these intermediate type positions place less importance on Inner-Directed behavior than do line managers but more importance than do staff managers, whereas for the Other-Directed traits they attach more importance than line managers but less importance than staff managers. The fact that respondents who report they have neither pure line jobs nor pure staff jobs are intermediate in their responses as well as in their organizational positions has two definite implications: The finding lends greater support to the finding of differences between the two extreme ends of the line-staff continuum; and, the finding also provides additional support to the argument advanced in an earlier paper on need satisfactions (Porter, 1963) that "apparently one can speak of a meaningful continuum going from pure line

³ The table presenting these mean values for each type of position within each management level has been deposited with the American Documentation Institute. Order Document No. 8026 from ADI Auxiliary Publications Project, Photoduplication Service, Library of Congress, Washington, D. C. 20540. Remit in advance \$1.25 for microfilm or \$1.25 for photocopies and make checks payable to: Chief, Photoduplication Service, Library of Congress.

jobs through jobs combining both line and staff features to pure staff jobs."

DISCUSSION

The previous article in this series that dealt with line versus staff differences in job attitudes showed that there were definite differences along this horizontal dimension of organization in terms of amount of perceived need satisfaction (Porter, 1963). The present article has also shown differences between the perceptions of line and staff managers, this time with regard to the relative importance attributed to certain personality traits for achieving job success. Considered together, the studies provide two independent pieces of evidence that the horizontal dimension of organization structure—i.e., line-staff distinctions—is a meaningful and useful variable to consider in making predictions about the job attitudes of managers. The differences between these two parts of the organization may be diminishing over time, as some observers believe, but they are still large enough to produce significant differences in managerial job attitudes.

The specific nature of the line-staff differences in job attitudes found in this study requires some discussion. The results showed a consistent trend across 4 of the 5 Other-Directed traits (with the fifth Other-Directed trait showing no trend) for staff managers to attach more importance to this type of behavior. It might be expected, then, that for the Inner-Directed traits line managers would consistently attach more importance than staff managers to this type of behavior. This did not happen in the sense that, although there was an overall trend in this direction, the trend did not hold consistently across all 5 Inner-Directed traits. Two of these Inner-Directed traits, Forceful and Decisive, did show very strong trends in the

expected direction with line managers attaching more importance. But, for 2 of the other Inner-Directed traits, Imaginative and Independent, there appeared to be clear (though not statistically significant) reversals with staff managers seeing them as more important compared to line managers. If the findings for all 10 traits—the 5 Inner-Directed and the 5 Other-Directed traits—are considered together, the conclusions would seem to be: Staff managers, as Whyte and others might predict, feel they have to show more Other-Directed behavior to succeed in their jobs than do line managers; but, they also apparently need to be more versatile than line managers since they indicate they have to emphasize certain aspects of Inner-Directed behavior as well as Other-Directed behavior. To achieve job success, line managers seemingly can concentrate on Inner-Directed type behavior, while staff managers may have to attend to aspects of both Inner- and Other-Directed behavior in order to be most successful in their staff jobs in American business organizations.

REFERENCES

- PORTER, L. W. Job attitudes in management: I. Perceived deficiencies in need fulfillment as a function of job level. *J. appl. Psychol.*, 1962, **46**, 375-384.
- PORTER, L. W. Job attitudes in management: III. Perceived deficiencies in need fulfillment as a function of line versus staff type of job. *J. appl. Psychol.*, 1963, **47**, 267-275.
- PORTER, L. W., & GHISELLI, E. E. The self perceptions of top and middle management personnel. *Personnel Psychol.*, 1957, **10**, 397-406.
- PORTER, L. W., & HENRY, MILDRED M. Job attitudes in management: V. Perceptions of the importance of certain personality traits as a function of job level. *J. appl. Psychol.*, 1964, **48**, 31-36.
- RIESMAN, D. *The lonely crowd*. New Haven: Yale Univer. Press, 1950.
- WHYTE, W. H., JR. *The organization man*. New York: Simon & Schuster, 1956.

(Received July 17, 1963)

A STUDY OF ATTITUDE CHANGE IN THE PRERETIREMENT PERIOD¹

SHOUKRY D. SALEH

Department of Civil Service, Ontario Provincial Government, Canada

2 separate sets of factors appear in preretirees' job attitude when they refer to their past experiences in middle age (30-55). Job-related factors provide satisfaction; context-related factors determine dissatisfaction. When sources of satisfaction were examined in the preretirement period, the dominant emphasis was on the context-related factors. This change of attitude was explained in view of the change of job structure. Choosing more attainable sources on the job, the context-related in case of preretirement, is more satisfying than choosing the ones which became more difficult to attain, the job-related factors.

The Motivation-Hygiene Theory presented by Herzberg, Mausner, and Snyderman (1959), was based on a study of engineers and accountants. The results showed that the job-related factors "motivators"² lead to positive job satisfaction while the context-related factors "hygienes"³ determine job dissatisfaction. It may be possible that these two groups stress the motivators more than the hygienes as the main source of satisfaction because of the nature of the engineering and accounting occupations. This also may be true because the respondents were mostly of middle age having many years left in their working careers in which they anticipate chances for advancement, growth in skill, more responsibility, and other motivators.

The purpose of this study is to test two hypotheses derived from the Motivation-Hygiene Theory on a different age group—preretirees of 60 to 65 years old. It is assumed that the job attitudes of older workers approaching retirement might be different from those of middle-age workers.

In this respect, Davis (1957) says that older workers tend to gradually develop into a separate social group, their interests and even their day-to-day conversations being

identifiably distinct. Turner's study (1955) showed that workers begin to feel hopeless as they contemplate growing older on a job which has little interest to them and which may in addition begin to demand too fast a pace considering their declining health and vigor. Moreover, in their review of research, Herzberg, Mausner, Peterson, & Copwell (1957) noted the importance of specific environment factors such as security and pay as the worker becomes older. Friedlander (1962) found that groups who derived satisfaction from the social and technical environment are frequently older than those who place prime importance on the more intrinsic job aspects which afford an opportunity for self-actualization.

In this view, two hypotheses were derived to be tested using a sample of preretirees.

1. Preretirees looking backward in their careers will indicate the motivators as the factors that give most satisfaction and the hygienes as the ones that determine dissatisfaction.

2. Preretirees looking toward the time left before retirement will indicate the hygienes as the important factors for job satisfaction.

METHOD

Subjects (Ss) were 85 male employees at managerial level. They were drawn from 12 Cleveland companies of different natures and sizes, all having a compulsory retirement plan at age 65. Their ages ranged from 60 to 65.

A semistructured interview used in the original Herzberg study was used to furnish the data for job satisfaction and dissatisfaction in middle age. The Ss were asked to relate critical incidents from

¹ This article is based upon a portion of a PhD dissertation at Western Reserve University, under the direction of Jay L. Otis.

² Achievement, recognition, advancement, growth in skill, responsibility, and interesting work.

³ Salary, interpersonal relations—supervisor, interpersonal relations—subordinates, interpersonal relations—peers, supervision, technical, company policy and administration, working conditions, factors in personal life, status, and job security.

TABLE 1

RELIABILITY OF THE INTERVIEW ANALYSIS

	Satisfaction		Dissatisfaction	
	First level	Second level	First level	Second level
Scorer 1	.97	.99	.86	.93
Scorer 2	.99	.97	.82	.33
Both scorers	.95	.99	.92	.80

their past experiences. To test the reliability of the interview analysis, two graduate students in psychology analyzed two-thirds of the total number of interviews. The correlations between the experimenter's analysis and the analysis of these two scorers are shown in Table 1.

The data for the sources of satisfaction in the pre-retirement period was obtained by a scale (Job Attitude Scale) consisting of 16 statements representing the six motivators and the 10 hygienes. Each statement was paired with the other 15 in a forced-choice format. The internal consistency (split half) of the scale was .94. This scale was given to the experimental group, age 60-65, as well as to a control group ($N = 39$), age 30-55.

RESULTS AND DISCUSSION

The results of the interview analysis supported the Motivation Hygiene Theory. They showed that when preretirees looked backward in their middle age, they indicated the motivators as the factors that provided satisfaction and the hygienes as the ones that determined dissatisfaction. Table 2 shows that the differences between the averages of the number of Ss who mentioned the motivators and those who mentioned the hygienes in the

TABLE 2

AVERAGE NUMBER OF SUBJECTS WHO MENTIONED THE MOTIVATORS (M) AND HYGIENES (H) IN THEIR PAST EXPERIENCES (MIDDLE AGE)

	Satisfaction sequence		Dissatisfaction sequence	
	First level	Second level	First level	Second level
Means of M ^a	17.5	22.5	5.8	3.3
Means of H ^b	1.6	5.2	7.1	6.1
Differences	15.9	17.2	1.3	2.8
<i>t</i> value	6.6	3.1	.56	2.5
<i>P</i> under 5 <i>df</i>	.01	.05	—	.05
<i>P</i> under 9 <i>df</i>	.001	.02	—	.05

^a Means of M, 6 factors.

^b Means of H, 10 factors.

sequences of satisfaction, are significant in both first and second level factors.⁴ The difference is also significant in the second level of the sequence of dissatisfaction. The motivators have the highest averages in case of satisfaction while the hygienes have the highest in case of dissatisfaction.

When the Ss indicated the sources of satisfaction in the preretirement period (60-65) on the "Job Attitude Scale," the picture was the reverse of that which they showed for their middle age (30-55), and the reverse of the picture shown by the control group (age 30-55). The experimental group indicated the hygienes in preretirement as the factors providing satisfaction while they stressed the

TABLE 3

MEANS OF MOTIVATORS (M) AND HYGIENES (H) INDICATED BY THE EXPERIMENTAL GROUP FOR THE PRERETIREMENT PERIOD AND BY THE CONTROL GROUP FOR MIDDLE-AGE PERIOD

	Experimental group	Control ^b group
Means of M	25.2	37.7
Means of H	34.7	22.2
Differences	9.5	15.5
<i>t</i> value	4.26	4.8
<i>P</i>	.001	.001

^a $N = 83$.

^b $N = 39$.

motivators in their middle age. The control group, who were middle aged, indicated the motivators as the source of satisfaction. Table 3 shows that the differences between the means of motivators and hygienes from the Job Attitude Scale are significant in both the experimental and control groups beyond the .001 level.

The results showed that the same Ss who indicated the motivators as the sources of satisfaction in their middle age indicated the hygienes in their preretirement years. This shift or change may have one or both of two explanations. One explanation is that the needs underlying the hygienes become more salient than the needs underlying the moti-

⁴ First level is the incident itself and second level is the subject's evaluation of it.

vators. The second one is that preretirees in the new circumstances related to their age, do not have access to the motivators so they choose the hygiene factors thinking that they can provide at least some satisfaction.

The first explanation stems from Maslow's hierarchy of needs, (1943, 1954). In this view, preretirees are not able to gratify the need for self-actualization as represented by the motivators because the needs that must be gratified first for this group are the safety need which is the basis of security, and the need for love which underlies interpersonal relations.

The assumption in the second explanation is that the goal of preretirees is the same as it was in middle age, namely, self-actualization. But because the chances to gratify this need through the motivators are not available, the preretirees shift to other goals despite the fact that these may not provide true satisfaction. According to the dissonance theory (Festinger & Aronson, 1960), if an individual is in a situation in which he continues to expend effort in order to reach some goal, yet does not reach it, he will experience dissonance. One way in which he could reduce dissonance would be by finding something about the situation to which he could attach value. Following this reasoning, when the preretirees were asked to choose between motivators and hygienes they did not pick the motivators which once provided them with satisfaction, simply because they realized that these factors became difficult to attain despite any expended effort. For instance, because they have only a few years left until retirement, companies usually do not give preretirees a chance for promotion, or for a new position which might give them feelings of growth. To reduce the dissonance aroused in this situation, preretirees choose the factors available or could be available—the hygienes. Because these factors do not provide intrinsic satisfaction, they may still feel uneasy in the

work situation. This might be one of the factors that the majority of the Ss expressed to the experimenter a favorable attitude toward their retirement.

The explanation based on the system of need hierarchy may lead to some misunderstanding. It assumes that the individual will not be able to actualize unless he has adequately satisfied all lower needs such as safety and love. This is not necessarily true because as implied in the Motivation Hygiene Theory and confirmed in this study, the channels for self-actualization and satisfaction are not on the same dimension with other needs which mainly have the power to reduce dissatisfaction. Thus, a person may be able to actualize while his needs for safety and love are not fully gratified. It appears then that the second explanation is to be preferred for understanding the shift in sources of satisfaction from the motivators in middle age to the hygienes in the preretirement period.

REFERENCES

- DAVIS, K. *Human relations in business*. New York: McGraw-Hill, 1957.
- FESTINGER, L., & ARONSON, E. The arousal and reduction of dissonance in social contexts. In D. Cartwright & A. Zander (Eds.), *Group dynamics*. Evanston, Ill.: Row, Peterson, 1960.
- FRIEDLANDER, F. *An analysis of the relationships among sources of job satisfaction*. Unpublished doctoral dissertation, Western Reserve University, 1962.
- HERZBERG, F., MAUSNER, B., PETERSON, E., & COPWELL, D. F. *Job attitudes: Review of research and opinion*. Pittsburgh: Psychological Service of Pittsburgh, 1957.
- HERZBERG, F., MAUSNER, B., & SNYDERMAN, B. *The motivation to work*. New York: Wiley, 1959.
- MASLOW, A. H. A theory of motivation. *Psychol. Rev.*, 1943, 50, 370-396.
- MASLOW, A. H. *Motivation and personality*. New York: Harper, 1954.
- TURNER, A. N. The older worker: New light on employment and retirement problems. *Personnel*, 1955, 32, 246-257.

(Received August 20, 1963)

CONVERGENT AND DISCRIMINANT VALIDITY FOR AREAS AND METHODS OF RATING JOB SATISFACTION¹

EDWIN A. LOCKE, PATRICIA CAIN SMITH, LORNE M. KENDALL,²
CHARLES L. HULIN,³ AND ANNE M. MILLER

Cornell University

A study to determine the convergent and discriminant validity of 4 rating methods and 5 areas of job satisfaction. Measures were administered to 133 randomly selected employees from 2 companies. A rating method employing a series of 6 faces ranging from a scowl to a smile and a direct graphic rating method were best according to the criteria of convergent and discriminant validity. All areas adequately satisfied both criteria, but the pay, promotions, and supervision areas showed somewhat greater discriminant validity than the work and people areas. The greater appropriateness of the convergent and discriminant criteria, as compared to other possible criteria, for demonstrating the validity of areas and measures of job satisfaction is discussed.

The problem of job satisfaction and its correlates has been of concern to both social and industrial psychologists for several decades. Despite a great deal of work in this area (for one review, see Herzberg, Mausner, Peterson, & Copwell, 1957) the relationships of satisfactions to other variables are far from clear. Findings previously accepted by many as fact are now being called into question (e.g., Hulin & Smith, 1963; Smith & Kendall, 1963b). In other areas of job satisfaction, findings have been so conflicting and equivocal that there is no semblance of a general law (cf. Brayfield & Crockett, 1955).

Differences in the methods employed by investigators is undoubtedly one of the numerous reasons for the differences in results obtained in various studies. Measures of satisfaction have often been chosen a priori and not subjected to rigorous validation procedures, or "validated" by procedures in-

appropriate to the measures used. Even when factor-analytic procedures have been employed, the items used were usually highly similar, with identical response formats, casting doubt upon whether the agreement of measures was attributable to method or to content.

We propose that the validity of measures of job satisfaction should be established by answering two important questions. What particular areas of satisfaction can be reliably discriminated by respondents? This is the question of the discriminant validity of different areas. What particular methods of measuring job satisfaction are most adequate and meaningful? This is the problem of the convergent validity of different methods. We substitute these questions for the more traditional one which would evaluate validity on the basis of the behavior which areas and methods would predict.

Industrial psychologists who have learned well the lesson that selection tests must be proved to be valid in predicting behavior have accepted the same paradigm for the validation of measures of satisfaction, and have attempted to use "objective" behavioral measures as criteria. This procedure faces both practical and logical objections.

Adequate behavioral criteria are simply not available for some areas (e.g., pay satisfaction). Where behavioral criteria are available and have been used, the results have

¹ These studies are a part of the Cornell University Studies of Retirement Policies, financed by a grant from the Ford Foundation. The present summary of research was supported in part by the Foundation for Research on Human Behavior. We wish to express our gratitude to the cooperating companies who made their records available and contributed the time of their personnel, and to the interviewers who contributed their time to make these studies possible.

² Now at Educational Testing Service, Princeton, New Jersey.

³ Now at the University of Illinois.

usually not been consistent from study to study (see Brayfield & Crockett, 1955), and even within the same company, different behavioral criteria do not agree (Seashore, Indik, & Georgopoulos, 1960).

Fortunately, it has occurred to several recent investigators that perhaps the theoretical basis for such an approach to validation is unsound. Katzell, Barrett, and Parker (1961), for example, have suggested that job satisfaction is best viewed as an output (dependent) variable rather than as an input (independent) variable. Our own position is similar to this. Job satisfaction is viewed as an affective response which is a result of experience on the job and which will function as an independent variable (that is, will affect behavior directly) only under very special circumstances, related to the individual and to his situation. (For a more complete account of our theoretical model, see Hulin, Smith, Kendall, & Locke, 1963; Smith, 1963; Smith & Kendall, 1963a.)

This point of view clearly makes the relationship of measures of satisfaction to behavior irrelevant to the validity of the measures. Validity must be demonstrated, instead, in terms of a relationship to other measures of the same satisfactions.

The model of convergent and discriminant validity proposed recently by Campbell and Fiske (1959) provides a method for this type of validation procedure. This model is applicable when each of several traits (or job areas) is measured by each of several methods. Briefly, the model requires that three basic criteria be fulfilled if a matrix of inter-correlations between each of several traits or areas measured by each of several methods is to show convergent and discriminant validity.

The demonstration of convergent validity requires that the correlations between the same areas as measured by different methods be "significantly different from zero and sufficiently large to encourage further examination of validity" (Campbell & Fiske, 1959, p. 82). Different methods of measuring the same trait or area should, therefore, agree.

The demonstration of discriminant validity involves two criteria. First, two different methods of measuring a given area should agree more closely with each other than with

any other method of measuring any other area. The correlations of the scores obtained by any two methods of measuring the same area should be higher than the correlations between scores for that area and scores for any other area using any other method. Secondly, the correlation between scores obtained by any two methods of measuring the same area should be higher than the correlations of scores in that area with scores in any other areas measured by the *same* method. In other words, the correlations within the same area, across different methods, should exceed the correlations between different areas within the same method.

The same criteria can be used to evaluate the relative effectiveness of different methods, as well as the discriminability of different areas by comparing the correlations obtained using each method.

The present study was designed to test the convergent and discriminant validity of five analytically distinguishable areas of job satisfaction, and of four varieties of graphic rating methods, which demanded little verbal sophistication of the respondents. This study was one of a series of pilot studies for a much larger study (see Smith & Kendall, 1963b) of satisfaction conducted in a nationwide sample of firms. Our purposes at this stage were to see if the five areas we had chosen, on the basis of an examination of the relevant literature, would fulfill the criteria of convergent and discriminant validity, and to find the nonverbal rating method which showed most convergent validity in the sense that it best represented other measures of the graphic type.

We intended to compare this rating method, for each area found to be discriminable, to a cumulated point scale of the check-list type (Guilford, 1954, p. 271) developed especially for this study. (For description and validation of the Job Descriptive Index see Kendall, Smith, Hulin, & Locke, 1963; Locke, Smith, Hulin, & Kendall, 1963).

It was our plan, therefore, to find the best possible graphic method in order that it could be compared with other, nongraphic, methods, rather than to use the one best rating measure as the only measure of satisfaction.

The present study not only yields substan-

TABLE 1

CORRELATIONS BETWEEN DIFFERENT METHODS OF RATING SATISFACTIONS WITH DIFFERENT AREAS OF JOBS FOR A RANDOM SAMPLE OF EMPLOYEES OF A FARMERS' COOPERATIVE

Scale	Faces					Graphic Direct					Graphic Triad					Boxes				
	A ₁	B ₁	C ₁	D ₁	E ₁	A ₂	B ₂	C ₂	D ₂	E ₂	A ₃	B ₃	C ₃	D ₃	E ₃	A ₄	B ₄	C ₄	D ₄	E ₄
Faces																				
Work	A ₁																			
Pay	B ₁	32																		
Promotion	C ₁	34	52																	
Supervision	D ₁	42	46	57																
People	E	51	31	28	44															
Graphic Direct																				
Work	A ₂	40	27	22	20	19														
Pay	B ₂	04	59	48	41	27	09													
Promotion	C ₂	19	17	55	28	19	42	45												
Supervision	D ₂	02	22	35	54	07	29	38	34											
People	E ₂	29	08	18	19	42	37	29	24	17										
Graphic Triad																				
Work	A ₃	37	09	25	19	15	62	08	26	14	31									
Pay	B ₃	12	47	46	30	23	08	66	33	22	22	06								
Promotion	C ₃	22	24	58	31	11	43	30	67	32	15	42	36							
Supervision	D ₃	02	12	27	38	10	14	26	27	50	09	30	30	41						
People	E ₃	30	14	16	07	25	24	20	15	01	57	40	30	25	17					
Boxes																				
Work	A ₄	43	26	13	16	22	47	17	16	08	23	26	08	12	01	13				
Pay	B ₄	19	46	17	10	22	19	43	14	07	21	05	23	11	02	06	63			
Promotion	C ₄																			
Supervision	D ₄	08	10	26	39	03	26	18	26	46	09	01	06	21	12	06	29	27		
People	E ₄	35	08	24	18	17	40	10	22	18	18	26	01	16	03	04	39	38		38

Note.—N = 41 male and 40 female combined; decimals omitted.

tive findings with respect to different areas and methods of measuring job satisfaction, but also illustrates what we feel to be a fruitful approach to the problems of validation in any situation in which the use of external or behavioral criteria is either illogical or impractical.

METHOD

The Subjects (Ss) in this study were 41 male and 40 female employees randomly selected from the payroll lists of a farmers' cooperative in New York State (Sample 1), and 52 male employees similarly randomly selected from a chemical company in Tennessee (Sample 2). These companies were selected because of the diversity of jobs and individuals represented. Mean annual earnings for males were, however, above community averages in both cases.

The job areas chosen for this study included work, pay, promotions, supervision, and people. Three types of scales were given to all employees in both samples; a fourth was used only in the first sample.

Faces Scales. Each employee completed a booklet containing five pages (one for each job area) each with six faces arranged across the page. The faces

ranged in six steps from a deep scowl to a broad smile.⁴ The Ss checked the face which showed how satisfied they were with each job area. Several different orders were used in arranging the pages. This format was chosen not only because it required negligible verbal ability on the part of the respondents, but also because it furnished a clear neutral point or zero point for each scale.

Direct Graphic Scales. In each questionnaire was a series of horizontal graphic scales with the end points labeled "0% satisfied" and "100% satisfied," and 10 unnumbered equal intervals marked off between. The Ss were asked to mark the point along the line which showed how satisfied they were with the job area in question.

Triadic Graphic scales. In addition to each of the graphic scales asking for ratings of the Ss' present jobs, there were two additional graphic scales. The first asked S to indicate how satisfied he would be with each area of the "worst" job he could imagine himself doing, and the second how satisfied he would be with that area of the "best" job he could imagine himself doing.

This triadic score for each area was based on the difference between the position checked for the

⁴ Appreciation is expressed to General Motors Corporation for permission to use the Faces scales.

TABLE 2

CORRELATIONS BETWEEN DIFFERENT METHODS OF RATING SATISFACTIONS WITH DIFFERENT AREAS OF JOBS FOR A RANDOM SAMPLE OF EMPLOYEES OF A CHEMICAL COMPANY

Scale		Faces					Graphic Direct					Graphic Triad				
		A ₁	B ₁	C ₁	D ₁	E ₁	A ₂	B ₂	C ₂	D ₂	E ₂	A ₃	B ₃	C ₃	D ₃ E ₃	
Faces																
Work	A ₁	<div></div>														
Pay	B ₁															
Promotion	C ₁															
Supervision	D ₁															
People	E ₁															
Graphic Direct																
Work	A ₂	33	11	22	06	25	<div></div>									
Pay	B ₂	35	31	48	28	00										
Promotion	C ₂	22	20	50	25	05										
Supervision	D ₂	53	39	45	50	08										
People	E ₂	17	10	16	07	43										
Graphic Triad																
Work	A ₃	07	03	03	04	20	29	14	22	29	03	<div></div>				
Pay	B ₃	25	12	34	26	02	22	76	53	45	24					
Promotion	C ₃	26	20	44	36	13	17	48	73	48	16					
Supervision	D ₃	33	27	27	34	02	27	45	24	75	00					
People	E ₃	05	04	07	03	24	05	00	10	21	33					

Note.—N = 52 males; decimals omitted.

“present” job and that checked for the worst job, in relation to the total distance between the checks for best and worst. This technique was explicitly designed to take into account the frame of reference of the individual.

Boxes Scales. Each employee in the sample from the farmers’ cooperative was interviewed, usually several days after the administration of the questionnaires. Each was shown drawings of rectangular boxes, one for each job area (with the exception of promotions). He was asked to imagine that all his feelings about a particular job area, both favorable and unfavorable, were “inside” the box, and then divide the area of the box into two sections according to what percentage of those feelings were satisfied and what percentage dissatisfied. The box was marked off in 10 unnumbered equal intervals along one of the longer sides. (Systematic bias was reduced by using nine different interviewers for obtaining these ratings. The interviewers did not, of course, see the questionnaires prior to the interview.)

PROCEDURE

All measures were intercorrelated separately for each company. The matrices are shown in Tables 1 and 2. The solid triangles show the intercorrelations of different areas (“traits”) measured by a single method; these will be called *heterotrait-monomethod*

triangles following the terminology of Campbell and Fiske (1959). The dotted triangles show the intercorrelations of different traits measured by different methods; these will be called the *heterotrait-heteromethod triangles*. The numbers in the diagonals of these heterotrait-heteromethod (dotted) blocks are the intercorrelations of the *same* traits measured by *different* methods; these will be called the *validity diagonals*, and the individual correlations, the *validities*. Unfortunately, no reliability estimates could be made of any of the measures, because testing time at this point seemed better spent in obtaining a broad range of ratings; the reliability diagonals are thus left blank.

RESULTS AND DISCUSSION

Areas

Demonstration of convergent validity for areas requires that the correlations in the validity diagonals be significantly different from zero. In Tables 1 and 2, 35 of 42 validity coefficients were significant at the .05 level

TABLE 3

VALIDITIES AVERAGED ACROSS TWO SAMPLES AND FOUR METHODS OF RATING SATISFACTIONS

Area	Average validity ^a
Work	.37*
Pay	.47*
Promotion	.59*
Supervision	.46*
People	.30*

^a Coefficients averaged after *z* transformation although the estimates from the two samples were given equal weight in averaging.

* $p \leq .05$ for $N = 133$.

or better. The average validities for all estimates for each area are shown in Table 3. All of these mean correlations were significant at the .05 level or better. Thus it appears that the matrices demonstrated adequate, although not perfect, convergent validity.

For discriminant validity, the first criterion is that each correlation in the validity diagonal be larger than the correlations in the same row and column in all the heterotrait-heteromethod (dotted) triangles. (This assumes a reflected matrix.) Let us count each time that a validity coefficient exceeds any single correlation in its row and column in a dotted triangle as a "correct heterotrait-heteromethod prediction." Table 4 shows the proportion of total correct predictions (for all methods combined) obtained in the two samples together. If we assume that half of

TABLE 4

PROPORTION OF CORRECT PREDICTIONS BY AREA BASED ON FIRST CRITERION FOR DISCRIMINANT VALIDITY FOR EACH COMBINATION OF METHODS

Area	Combined results for all methods in both samples ^a
Work	.90*
Pay	.89*
Promotion	.96**
Supervision	.92*
People	.74*

Note.—Proportion of validity coefficients larger than heterotrait-heteromethod correlations in same row and column.

^a $N = 133$.

* $p \leq .05$ for 183 predictions.

** $p \leq .01$ for 114 predictions.

the predictions would be correct by chance, then the total proportions in Table 4 are much larger than chance expectancy. The data, therefore, demonstrated adequate discriminant validity by the first criterion.

The second criterion for discriminant validity is that each correlation in the validity diagonal be larger than the correlation in the same row and column of the heterotrait-monomethod (solid) triangles (again assuming a reflected matrix). A "correct monomethod prediction" will occur when a validity exceeds a value in its row and column in a monomethod triangle. Table 5 shows the proportion of total correct monomethod predictions for each job area obtained in the two samples. For Pay, Promotions, and Supervision these proportions of correct predictions were significantly greater than the expected proportion of .50. The proportion of correct predictions for Work was not significantly greater than .50 and that for People was actually less than .50. This most rigorous

TABLE 5

PROPORTION OF CORRECT PREDICTIONS BY AREA BASED ON SECOND CRITERION FOR DISCRIMINANT VALIDITY FOR ALL METHODS IN BOTH SAMPLES

Area	Proportion of correct predictions
Work	.55
Pay	.65*
Promotion	.92*
Supervision	.62*
People	.46

Note.—Proportion of validity coefficients larger than heterotrait-monomethod correlations in same row and column.

* $p < .05$.

criterion of discriminant validity, therefore, was only partially met. A large proportion of the failures are contributed by the Boxes scales. Moreover, Campbell and Fiske (1959, p. 83), and Humphreys (1960) suggest that this second criterion is seldom met in research in individual differences; the results, therefore, are not discouraging.

Comparing the validities of the different areas, Pay, Promotions, and Supervision show the greatest convergent and discriminant va-

lidity, and Work and People the least. However, even these two areas fulfilled two of the three criteria adequately, indicating that they were to some extent discriminable. Thus it was decided to retain all areas for our later studies.

Methods

In order to compare the validity of the different methods, it was necessary to compute indices comparable to those in Tables 3, 4, and 5 for each method, (rather than for each area). Convergent validities as shown in Table 6 are not greatly or significantly different, but favor the Direct Graphic and the Faces methods.

The proportion of total correct heterotrait-heteromethod predictions, and the proportion of total correct monomethod predictions is given for each method in the two samples in Table 7.

The values in both tables are corrected by removing all predictions based on Direct Graphic—Triad Graphic diagonals, which, in addition to the spurious effects resulting from use of the same (Direct Graphic) score in both methods, may have been biased to some extent because the ratings of present, best, and worst jobs were made consecutively. In other words, the methods were not “maximally different” as Campbell and Fiske suggest should be the case. The correlations are probably a combination of reliability and validity estimates (Campbell & Fiske, 1959, pp. 83 ff.).

Comparing the methods shows greatest discriminant validity for the Direct Graphic method. The Faces method ran a very close second, and the Triad Graphic method came in last (cf. Table 7).

TABLE 6
CONVERGENT VALIDITIES OF METHODS AVERAGED
ACROSS FIVE JOB AREAS AND TWO SAMPLES

Method	Average validity ^a
Faces	.40
Direct Graphic ^b	.44
Triad Graphic ^b	.29
Boxes	.31

^a Averaged after *z* transformation.
^b Does not include Direct Graphic—Triad Graphic validities.

TABLE 7
DISCRIMINANT VALIDITIES OF FOUR RATING METHODS
BASED ON COMBINED RESULTS FOR FIVE JOB AREAS

Method	Proportion of heterotrait-heteromethod values smaller than validity in same row and column	Proportion of heterotrait-monomethod values smaller than validity in same row and column
Faces	.89	.55
Direct Graphic ^a	.91	.69
Triad Graphic ^a	.77	.35
Boxes	.77	.39

^a Does not include Direct Graphic—Triad Graphic correlations.

Before deciding upon a method of rating, the distributions of the different methods were also compared. The Faces method gave wider and more nearly unimodal distributions, and was finally chosen as the more satisfactory of the rating methods.

Two points should be kept in mind when evaluating the present results. Although the various rating methods showed considerable agreement with each other when measuring each of several traits, this does not mean necessarily that the measures are all equivalent, or even that those showing the most agreement are equivalent, in the sense of having similar correlations with other measures differing in the operation of measurement or in the traits measured. Although Faces and Direct Graphic methods showed considerable agreement in terms of convergent and discriminant validity, to treat them as equivalent in terms of expected correlations with other measures could prove misleading (Smith & Kendall, 1963b). Lack of equivalence is a great inconvenience, but must be faced when selection must be made from among several methods of measuring an area. If all rating alternatives cannot be used, then the investigator must settle upon whatever method, after consideration of convergent and discriminant validity, appears to represent what is common to all the methods.

The present study, moreover, demonstrates only that the Faces method was the best graphic rating method of the four employed.

This does not mean that it would necessarily show convergent and discriminant validity when compared to maximally different methods of measuring satisfaction. Similarly, there is no guarantee that the five areas measured would be as discriminable when maximally different methods of measuring them were used.

In spite of the possible difficulties with such a procedure, we feel that the Campbell-Fiske model provides a better method of determining the adequacy of a number of different methods and areas than either a priori choice of method and areas, or an attempt to validate by correlating with behavioral criteria. Predictive validity, then, though an essential characteristic of many psychological measures, can be inappropriate for testing the adequacy of other measures. Since job satisfaction inventories are practically never used as selection devices, predictive validity is not really a relevant consideration here. In addition, if we accept the fact that satisfaction is an output variable then we eliminate the necessity of demonstrating any kind of predictive validity for such measures in order to declare them useful and meaningful. This should not be taken, however, as license for using whatever method the investigator wishes. Rather, other rigorous criteria such as those required by the Campbell-Fiske model are necessary in this situation. In some respects this model provides better criteria than does the requirement of predictive validity, since the latter can be obtained without the investigator actually knowing what his measure is measuring. The Campbell-Fiske model provides for a rigorous kind of "construct" validity, which, if not a substitute for predictive validity, is for certain measures, a superior replacement for it.

REFERENCES

- BRAYFIELD, A. H., & CROCKETT, W. H. Employee attitudes and employee performance. *Psychol. Bull.*, 1955, **52**, 396-424.
- CAMPBELL, D. T., & FISKE, D. W. Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychol. Bull.*, 1959, **56**, 81-105.
- GUILFORD, J. P. *Psychometric methods*. New York: McGraw-Hill, 1954.
- HERZBERG, F., MAUSNER, B., PETERSON, R. O., & COPWELL, DORA F. *Job attitudes: Review of research and opinion*. Pittsburgh: Psychological Service of Pittsburgh, 1957.
- HULIN, C. L., & SMITH, PATRICIA C. Sex differences in job satisfaction. *J. appl. Psychol.*, 1964, **48**, 88-92.
- HULIN, C. L., SMITH, PATRICIA C., KENDALL, L. M., & LOCKE, E. A. Cornell studies of job satisfaction: II. Model and method of measuring job satisfaction. Ithaca: Cornell University, 1963. (Mimeo)
- HUMPHREYS, L. G. Note on the multitrait-multimethod matrix. *Psychol. Bull.*, 1960, **57**, 86-88.
- KATZELL, R. A., BARRETT, R. S., & PARKER, T. C. Job satisfaction, job performance, and situational characteristics. *J. appl. Psychol.*, 1961, **45**, 65-72.
- KENDALL, L. M., SMITH, PATRICIA C., HULIN, C. L., & LOCKE, E. A. Cornell studies of job satisfaction: IV. The relative validity of the job descriptive index and other methods of measurement of job satisfaction. Ithaca: Cornell University, 1963. (Mimeo)
- LOCKE, E. A., SMITH, PATRICIA C., HULIN, C. L., & KENDALL, L. M. Cornell studies of job satisfaction: V. Scale characteristics of the job descriptive index. Ithaca: Cornell University, 1963. (Mimeo)
- SEASHORE, S. E., INDIK, B. P., & GEORGOPOULOS, B. S. Relationships among criteria of job performance. *J. appl. Psychol.*, 1960, **44**, 195-202.
- SMITH, PATRICIA C. Cornell studies of job satisfaction: I. Strategy for the development of a general theory of job satisfaction. Ithaca: Cornell University, 1963. (Mimeo)
- SMITH, PATRICIA C., & KENDALL, L. M. Cornell studies of job satisfaction: VI. Implications for the future. Ithaca: Cornell University, 1963. (a) (Mimeo)
- SMITH, PATRICIA C., & KENDALL, L. M. Achieving generality in studies of satisfaction. Ithaca: Cornell University, 1963. (b) (Mimeo)

(Received August 21, 1963)

RELATIONSHIPS BETWEEN JOB DIFFICULTY, EMPLOYEE'S ATTITUDE TOWARD HIS JOB, AND SUPERVISORY RATINGS OF THE EMPLOYEE EFFECTIVENESS

BYRON SVETLIK, ERICH PRIEN, AND GERALD BARRETT

Psychological Research Services, Western Reserve University

Using correlational techniques, an investigation was made of the relationships between job difficulty (estimated by job evaluation factors), employee attitudes toward his job, and job environment, and supervisory ratings of employee performance. As job difficulty increased, employee attitudes were significantly more positive toward the job, management, and communication, and opportunity for advancement. Partial correlations showed that the relationship between job satisfaction and job difficulty increased when the effects of general morale were eliminated. Supervisory ratings of employee effectiveness were significantly rated (negatively) to employee salary and job tenure. Correlations between employee's attitude dimensions indicate increasing complexity of job content and increased content with people as a part of the job, are positively related to an employee's attitude toward his job.

Herzberg (Herzberg, Mausner, & Snyderman, 1959) has suggested that the critical determinants of how a man feels about his job are lodged in the intrinsic characteristics of the job itself, not in the environmental characteristics surrounding the job. Herzberg interprets his findings based on work with professional and high level employees as a verification of the importance of the job itself as a determinant of job satisfaction. There has been some criticism that these findings are not generalizable and consequently will not hold up for employees holding lower level jobs. Friedman and Havighurst (1954) point out that for individuals gainfully employed, including those with low skill levels as well as professionals, work had meaning in addition to that of earning a living. While the design of their study did not permit statistical proof they interpreted the trend of the data as meaning—skilled craft and white collar groups obtain meaning from the job on the bases of things other than monetary rewards. Harding and Naurath (1960) showed that an individual's ratings of his work activities are affected by his amount of work experience and the organization to which he belongs. Prien, Otis, Campbell, & Saleh (1963) and Saleh, Prien, Otis, & Campbell (1963) also showed the import-

ance of job level and of situational factors as determinants of job attitudes. Svetlik (1961) in unpublished research showed that job attitudes were related to personality characteristics and that respondents holding jobs had more positive attitudes than did individuals indicating some degree of vocational maladjustment. The indication was the seeking of vocational counseling. Obviously the relation is complex. A comprehensive theory of job satisfaction likely will include some reference to situational characteristics, job characteristics and personal characteristics of the workers.

The purpose of this paper is to investigate the relation of job attitudes to selected situational factors, personal characteristics of respondents, and job characteristics. While the research is set in a single multidepartment company, the range of control variables is spanned. Data were collected by members of Psychological Research Services of Western Reserve University during the process of developing and installing a wage and salary program for all nonunion employees.

METHOD

Procedure

Three types of data were collected for investigation for this research. Company employees were

interviewed to obtain a description of their job. Supervisors were requested to complete a performance evaluation of each of their subordinates. And each employee was asked to rate his own job attitudes. The questionnaires were completed department by department with an examiner, who was a member of the consulting organization, present to answer employee questions and to assure and guarantee anonymity of their responses as far as the company was concerned. Employees signed their returns so that this analysis could be completed. They were told the data would be used for research purposes only, and that no individual would be identified. The interviewers had developed good rapport with the employees during the time the job-analysis information was collected, and observed that

the employees seemed quite willing to be helpful and candid in their responses.

The two questionnaires, one to measure the employees' job attitude, and the other for the supervisors' ratings of subordinates' performance, were developed especially for this research project. Both forms use a graphic scale format. Dimensions included in the attitude-measurement form are those which have been repeatedly identified by factor analyzing multiple question attitude questionnaires. The Employing Rating of Various Job Factors asked the employee to express his feelings on the following scales: adequacy of supervision, job satisfaction, management-communications, rewards, advancement, and general morale. Below is an example of one of the scales.

JOB SATISFACTION: What are your present feelings toward your job. The role the job itself occupies in your life. Consider your interest in your job future, pride, and self-evidence of job suitability.

VERY LOW	BELOW AVERAGE	AVERAGE	ABOVE AVERAGE	HIGH
-------------	------------------	---------	------------------	------

The Supervisor Rating of Subordinates Performance asked the supervisor to rate his employees on four dimensions of performance—personal relations, job competence, job energy, and overall effectiveness. Following is an example of one of these scales.

JOB COMPETENCE: The effectiveness with which this person plans, solves problems, allocates work, and knows and applies technical knowledge to his job.

INEFFECTIVE	SOME SHORT- COMINGS	AVERAGE	FEW SHORT- COMINGS	EXTREMELY EFFECTIVE
-------------	------------------------	---------	-----------------------	------------------------

The personnel office of the company provided the following information: employees' salaries, tenure on the job, and company tenure.

In the development of the wage and salary plan, each job was rated using the Personnel Research Institute's Job Evaluation Manual for Clerical and Administrative positions. The manual provides a scaling of jobs on 12 factors. Each level of each factor is assigned a relative point value based on judged proportionate importance of the factor. The point levels for each of the 12 factors were totaled to give the total job difficulty. This served as the estimate of job difficulty used in this study. Harding, Madden, and Coleson (1960) have shown that it is not advisable to use an abbreviated job-evaluation plan. Consequently, all 12 factors were used. While less convenient, the increase in reliability was desirable.

The relative salary was derived by computing a least squared regression line for all the jobs for the relationship between job difficulty and salary. The relative salary was graphically derived and is the difference between the actual salary and the predicted salary.

Pearson correlations were computed between job level for each of the 12 factors; salary, total job-difficulty points, job tenure, company tenure, relative salary, employee job-attitude ratings, the supervisor ratings of employee job performance. Means and

standard deviations were computed for each of the variables. The data reported in this study are shown in Table 1. Selected partial correlations were computed and are reported in the text.

RESULTS AND DISCUSSION

The major purpose of the study was to investigate the relationship between job difficulty and the employees' attitudes. According to current theory, it was hypothesized that variation of job difficulty would be accompanied by concomitant variation in the employee attitudes. As illustrated by the results in Table 1, the correlations were in the expected direction and highly significant. The correlations were significant between job difficulty and employee's rating of his feelings toward the job (job satisfaction), timely notice and explanation of things that affect his job (management-communication), and his feelings toward a career with the company, training and promotion policies (advancement). As anticipated, the correlations were insignificant between job difficulty and the employee's attitude toward the supervisor's

TABLE 1

CORRELATIONS, MEANS, STANDARD DEVIATIONS BETWEEN EMPLOYEE ATTITUDE DIMENSIONS, SUPERVISOR RATINGS AND SELECTED JOB CHARACTERISTICS

Variable	Salary	Job diffi- culty points	Job tenure in years	Company tenure in years	Relative salary	<i>M</i>	<i>SD</i>
Attitude dimensions ^a							
Employee attitude toward supervision	06	−03	13	08	00	3	0.8
Job satisfaction	24*	22*	29**	27**	06	3	0.9
Management and communications	12	19*	06	15	02	3	0.9
Rewards	09	15	12	14	00	2	1.1
Advancement	19*	28**	−06	−02	06	3	1.0
General morale	02	00	32**	21*	01	3	0.9
Supervisor ratings ^b							
Personal relations	04	13	00	−11	−14	3	0.8
Job competence	12	17*	−05	−02	−01	3	0.7
Job energy	07	12	−03	−06	−20*	4	0.8
Overall effectiveness	−21*	00	−13	−27**	−17*	3	0.7
<i>M</i>	491.	275	6.3	13.4	−17		
<i>SD</i>	149	93.4	6.26	10.7	75.5		

^a *N* = 110.
^b *N* = 130.
* *p* < .05.
** *p* < .01.

skill in handling human relations and his job know-how (supervision), rewards and recognition, equitable pay, and status (rewards), and the employee's overall feeling of well being both on and off the job (general morale).
The most interesting result of this analysis is the lack of correlation between employee general morale and job difficulty ($r = .00$) and also with salary ($r = .02$). Yet the employee's general morale is significantly correlated with job and company tenure. The constellation of correlations suggests that employee general morale is colored by his environment or perhaps his personal characteristics in terms of his needs and expectations rather than his perception of his job. Perhaps the controlling factor is the employee's perception of the company as a place to work in the community. The high correlation between morale and job tenure could be explained in light of Harding and Naurath's (1960) findings. They show that the amount of work experience of an individual does affect his rating of the work activities that make up his job. The results here confirm the positive

relation between tenure and employee job attitudes. A longitudinal study might clarify the nature of the relationship, as being a function of selective attrition or an actual change in attitudes as suggested by Harding and Naurath. A most important point that bears repeating is—the employee's general morale is not related to the level of difficulty of his job. Morale seems to be a function of the person's life situation rather than his job although it has its impact in the job situation as evidenced by job tenure.
Job difficulty has been hypothesized to be an important variable in explaining job satisfaction. Selective attrition also may play a part in affecting job satisfaction, and job morale, and the relationship between these variables. The correlation between job satisfaction and job difficulty with the effects of job tenure eliminated, were not significantly different ($r = 12.3 = .21$). Likewise the correlations did not change when general morale and job tenure were correlated eliminating the effects of job difficulty ($r = .32$ in both cases).

However, the correlation between job satisfaction and job difficulty rose from .22 to .29 when the effects of general morale were partialled out. Percentage of explained variance increased from 4 to 8%. This result appears to be more evidence for theorists who hold that an individual's job satisfaction is influenced by the intrinsic characteristic of the job.

Most of the correlations between the supervisory ratings of employee effectiveness and situational factors were insignificant. The exception was the negative correlations between the supervisor ratings of the employee's overall effectiveness and the employee's salary and job tenure. This suggests that the supervisor is more favorably impressed with the newer employee. Employees of this company, however, tend to stay with the organization (mean tenure = 10.7 years). Again, the exception in this organization is that there is a high turnover rate for the supervisory positions due to both promotion and termination. The relation between tenure and performance ratings in this is not startling since the new supervisor often brings his own subordinates into the work group, individuals

who have been with the organization for shorter periods of time and who are not yet at the salary level of the long-term employee. The design of this experiment did not provide information as to whether or not the supervisors' judgments were valid.

A further clarification of the relationships between job attitudes and situational or job characteristics was obtained by computing correlations between each of the employee's attitude dimensions and the job-evaluation factors score. To a degree, the job evaluation represents a measurement of job content in that factors and each level is defined in operational terms. That is, the factors represent a constant scoring system for the description of the job content. These results appear in Table 2.

As can be seen, the most obvious relationship is between attitude toward advancement and job level. The higher the job level, the more satisfied the person appears to be with the advancement potential in his job. The exception is with Variable 3—which is the measurement of dexterity required. In this case, as requirements for dexterity are increased, attitudes toward job satisfaction,

TABLE 2

CORRELATIONS BETWEEN FACTORS IN A JOB EVALUATION MANUAL AND EMPLOYEE ATTITUDES

Variable	Super- visor	Job satis- faction	Manage- ment-com- munication	Rewards	Advancement	Morale
Attitude dimensions ^a						
Work experience	-.01	.17	.06	.06	.16	-.03
Knowledge and training	.03	.08	-.01	.00	.15	-.12
Dexterity	-.08	-.21*	-.20*	-.14	-.23*	-.02
Supervision received	-.07	.16	.18	.13	.27**	.00
Supervision given	.04	.27**	.25**	.16	.33**	.09
Number supervised	-.09	.21*	.22*	.14	.33**	.04
Responsibility for money	-.15	.15	.05	.07	.13	.00
Responsibility for confidential information	.00	.04	.13	.11	.23*	-.07
Responsibility for getting along with others	-.02	.23*	.10	.15	.20*	-.01
Responsibility for accuracy	-.02	.18	.12	.13	.22*	-.02
Pressure of work	-.13	.16	.20*	.08	.14	.16
Unusual working conditions	.07	-.10	.01	.08	.06	-.06

^a $N = 110$.* $p < .05$.** $p < .01$.

management-communication, and advancement seem to decrease. In effect, this is quite reasonable in that jobs which require maximum dexterity are also those in which the individual has little or no contact with other people. These are jobs which require concentration on detail and very often involve working in isolation. Job satisfaction and management-communication are further related to the amount of supervision given and the number supervised. However, job satisfaction is related to responsibility for getting along with others, and attitude towards management-communications is related to the amount of pressure in the job. An overall observation is that employee attitudes appear to be a function of job content, at least to the extent that jobs vary in complexity and difficulty level. Again, the criticism noted earlier regarding selective attrition also applies here. The finding may merely reflect the fact that individuals who are satisfied with the company are those who are promoted and remain with the company, whereas those who are dissatisfied tend to leave or not be promoted or possibly discharged.

Also of some importance is the lack of correlation between attitudes and the dimension of work experience, knowledge and training, responsibility for money, and working conditions. These are primarily impersonal factors in the job as compared with the factors, supervision received, supervision given, number supervised, and responsibility for

getting along with others. These combinations of variables show considerable correlation so that one might speculate in regarding the importance of other people in the work situation. Again, the reversed correlation with the dexterity required support this approach. We conclude that increasing complexity of job content and increased contact with people as a part of the job are possibly related to an employee's attitude toward his job.

REFERENCES

- FREIDMAN, E. D., & HAVIGHURST, R. J. *The meaning of work and retirement*. Chicago: Univer. Chicago Press, 1954.
- HARDING, F. D., MADDEN, J. M., & COLSON, K. Analysis of a job evaluation system. *J. appl. Psychol.*, 1960, **44**, 354-357.
- HARDING, F. D., & NAURATH, D. A. Effects of job experience and organization on the rating of tasks. *Engng. Industr. Psychol.*, 1960, **2**, 63-68.
- HERZBERG, F., MAUSNER, B., & SNYDERMAN, P. D. *The Motivation to work*. New York: Wiley, 1959.
- PRIEN, E. P., OTIS, J. L., CAMPBELL, J. T., & SALEH, S. D. The measurement of job attitudes: A comparison of methods. Cleveland: Western Reserve University, Psychological Research Services, 1963. (Mimeo)
- SALEH, S. D., PRIEN, E. P., OTIS, J. L., & CAMPBELL, J. T. The measurement of job attitudes: Job attitudes, performance and job level. Cleveland: Western Reserve University, Psychological Research Services, 1963. (Mimeo)
- SVETLIK, B. L. A study of the relationship of expressed attitudes about one's job and job conditions to personality variables as measured by various personality tests. Unpublished master's thesis, Western Reserve University, 1961.

(Received August 23, 1963)

AN ANALYSIS OF VOCATIONAL INTERESTS AT TWO LEVELS OF MANAGEMENT¹

HRACH BEDROSIAN²

Teachers College, Columbia University

The differences in the vocational interests of top and middle management level personnel of a large, multiplant industrial corporation were studied. Each S was classified according to work level, field, and work role (line or staff). Top management Ss were found to have a higher socioeconomic level of vocational interest than middle management Ss. Clarity of interest patterning was not related to work role, nor, except in one case, was it related to managerial level of work. No differences were found in the decisiveness with which top and middle management Ss responded to interest-test items.

The psychological and sociological factors associated with attaining high occupational status have been the subject of intensive research for many years. While recognizing that any or all of these factors may be important in the achieving of managerial status, this study investigated the vocational interest characteristics of individuals who have achieved managerial status.

Information suggesting the nature of the interest characteristics of men in managerial positions was drawn from previous studies (Strong, 1927, 1946) of the interests of management personnel, by studying the managerial job itself (Barnard, 1948; Livingston, 1960; Shartle, 1956), and by inference from the personality characteristics of men in management jobs (Coates & Pellegrin, 1957; Ghiselli, 1959; Harrell, 1961; Henry, 1949).

Hypotheses

The following hypotheses were formulated:

1. Men in top management positions differ in vocational interests from men in middle management positions in the following ways: (a) The vocational interests of men in top management are less clearly patterned than those of men in middle management. (b) Men in top management have a higher socioeconomic level of vocational interest than men in middle management. (c) Men in top man-

agement are more decisive in their expressions of vocational interest than men in middle management.

2. Men in line managerial positions differ in their vocational interests from men in staff positions in the following way: The vocational interests of men in line managerial positions are less clearly patterned than those of men in staff positions.

PROCEDURE

Subjects

The subjects (Ss) were drawn from the management personnel of a large multiplant corporation engaged in the design, development, and manufacture of heavy industrial machinery. The corporation's divisions and plants were scattered throughout the United States. Collection of data was carried out as part of a personnel research program for which the corporation had engaged the services of an industrial consulting organization.³

The primary instrument used in this study was the Strong Vocational Interest Blank (SVIB) for Men. Analysis of the Strong profile was carried out by grouping the occupational scales as prescribed by Darley and Hagenah (1955, p. 83).

Each S was classified according to work level, field, and work role (line or staff).

Level of Work

The Ss were assigned to one of two managerial groups: top management or middle management. Top management was composed of the president of the corporation, the vice presidents, the general managers of operating divisions, and department managers in the operating divisions and corporate headquarters.

³This study was completed as part of studies conducted by Kenneth F. Herrold and Associates, New York, a nonprofit professional organization engaged in personnel research.

¹ The study is based on a doctoral dissertation submitted to Teachers College, Columbia University. The members of the dissertation committee were Donald E. Super, Kenneth F. Herrold, Jean P. Jordaan, and Albert S. Thompson.

² The author is now at New York University.

Middle management was composed of section heads and supervisory personnel from each of the divisions and corporate headquarters. Each middle management *S* was an immediate subordinate to a top management person.

Field of Work

Each *S* was assigned to a field of work category based on job title, departmental membership, and a detailed description of job duties, and functions. Using the system of occupational classification proposed by Moser, Dubin, and Shelsky (1956), *Ss* were assigned to one of three fields of work: business contact, business administration and control, and technology.

Work Role (Line or Staff)

In addition to distinguishing managerial jobs on the basis of level and field of work, a third distinction was made based on the degree of task specialization or scope of managerial responsibility. Line and staff were operationally defined as follows:

Line managers are those who direct a work unit, whose primary duties consist of the planning, organizing, controlling of programs and personnel supervision. Their place on the organization chart is in the vertical chain of command of the organizational hierarchy.

Staff specialists are those management level personnel who perform a narrow range of specialized tasks such as research, who act primarily as consultants or advisors to line managers, who do not supervise groups of subordinates (may have a secretary and/or an assistant reporting to him), or who provide special services to line managers.

Dependent Variables

Clarity of interest patterning refers to the degree of differentiation of interests and disinterests in the interest profile. Cronbach and Gleser (1953) have termed this profile characteristic "scatter." Scatter scores were computed by:

1. Combining the SVIB occupational scales to form the seven interest families as designated by Darley (Darley & Hagenah, 1955).
2. Deriving a mean score for each interest family by summing the standardized scores of each of the occupational scales in that family and dividing by the number of occupational scales in the interest family.
3. Deriving an overall mean score for all of the interest families by summing the mean scores of the interest families and dividing by the number of interest families. There are seven interest families.
4. Squaring each of the deviations of the interest family means from the mean of all the interest families, and summing these squared deviations.

The scatter score is the square root of the sum of the squared deviations.

Socioeconomic level of interest or interest level as defined by Barnett, Handelsman, Stewart, & Super (1952) was measured by the Occupational Level (OL) scale of the SVIB.

Decisiveness of expressions of vocational interest was measured by the relative absence of "indifferent" responses to Items 1 through 280 of the SVIB. The remaining items, 281 through 400, include forced choice, comparison, and self-rating items, and are therefore not suitable for inclusion. Fewer indifferent responses were taken to indicate more decisive expressions of interest.

RESULTS AND DISCUSSION

As shown in Table 1, only in the business contact group are the interest profiles of top management men less clearly patterned than men in middle management. The possibility that the business contact group was unique in some respect was considered. In reviewing the educational background of all *Ss*, it was found that 71% of the *Ss* in business administration had backgrounds in business and liberal arts, and 97% of the *Ss* in technology had backgrounds in engineering and the physical sciences. This is not surprising, since there are formal programs of study which prepare men for careers in the kinds of occupations which made up these fields of work.

There is, however, no formal program of study that prepares one for careers in the occupations which comprise the business contact field. Occupations in business contact work, therefore, are filled by individuals with a variety of educational backgrounds. For

TABLE 1
COMPARISON OF SCATTER SCORES BY MANAGERIAL LEVEL AND FIELD OF WORK

Field and level of work	<i>N</i>	<i>M</i>	<i>SD</i>	<i>t</i>
Business contact				
Top management	45	22.29	8.21	2.36*
Middle management	59	26.39	9.30	
Business administration				
Top management	50	21.82	7.17	1.21
Middle management	50	23.68	8.08	
Technology				
Top management	52	19.52	6.27	0.24
Middle management	103	19.64	5.88	

* *p* ≤ .05.

example, 76% of the business contact Ss were found to have educational backgrounds in engineering and physical sciences. Conceivably, generalized or undifferentiated interests may be more closely associated with high occupational level achievement in occupations where there is no formal program of study preparatory to work in the occupation than in occupations where formal programs of study are a virtual prerequisite.

Table 2 shows that the occupational level scores of men in top management are significantly higher than the scores of men in middle management in each of the three fields of work. Strong (1946) found differences in interest level between men in managerial and nonmanagerial positions. The present data

TABLE 2

COMPARISON OF OCCUPATIONAL LEVEL (OL) SCORES OF MANAGERS BY LEVEL AND FIELD OF WORK

Field and level of work	<i>N</i>	<i>M</i>	<i>SD</i>	<i>t</i>
Business contact				
Top management	45	59.36	7.97	3.37*
Middle management	59	54.85	4.52	
Business administration				
Top management	50	57.02	7.31	4.10*
Middle management	50	52.02	4.42	
Technology				
Top management	52	57.52	6.33	6.16*
Middle management	103	51.48	3.98	

* $p \leq .01$.

carry Strong's findings one step further, and suggest that there are differences in interest level among individuals within the managerial group itself.

Interests can be used to predict job satisfaction. Would an OL score be useful in predicting the managerial level at which an individual "is most likely to find appropriate outlets for his interests?" (Barnett et al., 1952). This question cannot be answered until more is known about the degree to which interests are acquired on the job.

Table 3 reveals that there were no differences in the frequency with which top and middle management Ss responded "indifferent" to SVIB items 1 through 280. As here defined, men in top management are no more decisive

TABLE 3

COMPARISON OF NUMBER OF "INDIFFERENT" RESPONSES MADE BY TOP AND MIDDLE MANAGEMENT Ss IN A GIVEN FIELD OF WORK

Field and level of work	<i>N</i>	Mean Number of "I" Responses	<i>SD</i>	<i>t</i>
Business contact				
Top management	45	107.11	24.44	1.81
Middle management	59	97.00	31.33	
Business administration				
Top management	50	105.80	28.16	0.02
Middle management	50	105.70	25.87	
Technology				
Top management	52	102.58	28.32	0.26
Middle management	103	103.84	28.50	

in their expressions of interest than men in middle management. It may be that their greater decisiveness is more manifest in business and social situations than in highly structured test situations such as a responding to SVIB items.

The hypothesis that men in staff positions have more clearly patterned interests than men in line positions, was not confirmed. These data are reported in Table 4.

While the clarity of interest patterns of men in line and staff positions does not differ, as suggested by the present data, there may be differences in the effectiveness with which these individuals are able to discharge their responsibilities and/or the degree of satisfaction which they derive from their work, since there is evidence (Burchard, 1954; Getzels & Guba, 1954) that job satisfaction and performance are impaired when men are forced to adopt work roles that conflict with their own work expectancies.

TABLE 4

COMPARISON OF SCATTER SCORES OF LINE AND STAFF Ss

Group	<i>N</i>	<i>M</i>	<i>SD</i>	<i>t</i>
Line managers	322	22.09	6.74	1.29
Staff managers	37	20.54	7.86	

REFERENCES

- BARNARD, C. I. *The functions of the executive*. Cambridge: Harvard Univer. Press, 1948.
- BARNETT, G. J., HANDELSMAN, I., STEWART, L. H., & SUPER, D. E. The occupational level scale as a measure of drive. *Psychol. Monogr.*, 1952, **66**(10, Whole No. 342).
- BURCHARD, W. H. Role conflicts of military chaplains. *Amer. sociol. Rev.*, 1954, **19**, 528-535.
- COATES, C. H., & PELLEGRIN, R. J. Executives and supervisors: Contrasting self perceptions and conceptions of each other. *Amer. sociol. Rev.*, 1957, **22**, 217-220.
- CRONBACH, L. J., & GLESER, G. C. Assessing similarity between profiles. *Psychol. Bull.*, 1953, **50**, 456-473.
- DARLEY, J. G., & HAGENAH, THEDA. *Vocational interest measurement: Theory and practice*. Minneapolis: Univer. Minnesota Press, 1955.
- GETZELS, J. W., & GUBA, E. G. Role, role conflict, and effectiveness. *Amer. sociol. Rev.*, 1954, **19**, 164-174.
- GHISELLI, E. E. Traits differentiating management personnel. *Personnel Psychol.*, 1959, **12**, 535-544.
- HARRELL, T. W. *Managers' performance and personality*. Cincinnati: South-Western, 1961.
- HENRY, W. E. The business executive: A study of the psychodynamics of a social role. *Amer. J. Sociol.*, 1949, **54**, 286-291.
- LIVINGSTON, R. T. *The manager's job*. New York: Columbia Univer. Press, 1960.
- MOSER, HELEN P., DUBIN, W., & SHELSKY, I. M. A proposed modification of the Roe Occupational Classification. *J. counsel. Psychol.*, 1956, **3**, 27-31.
- SHARTLE, C. L. *Executive performance and leadership*. Englewood Cliffs, N. J.: Prentice-Hall, 1956.
- STRONG, E. K., JR. Interests of senior and junior public administrators. *J. appl. Psychol.*, 1946, **30**, 55-71.
- STRONG, E. K., JR. Vocational guidance of executives. *J. appl. Psychol.*, 1927, **11**, 331-347.

(Received August 26, 1963)

THE ANALYSIS OF JOB PERFORMANCE BY MULTIDIMENSIONAL SCALING TECHNIQUES¹

DOUGLAS G. SCHULTZ AND ARTHUR I. SIEGEL

Applied Psychological Services, Wayne, Pennsylvania

The application of multidimensional scaling methods to the analysis of job performance was explored. Men experienced in the job of Naval aviation electronics technician designated 18 tasks as constituting that job at the entry levels. Supervisory personnel then judged the similarity between all pairs of these tasks. The resulting scaled similarity estimates were analyzed by standard multidimensional scaling techniques. The work performed by aviation electronics technicians at the job entry levels was perceived by supervisors as involving 4 basic dimensions. It appears to be feasible and fruitful to apply multidimensional scaling techniques to the analysis of job performance.

One of the more evident facts about job performance criteria is that characteristically they are complex and multidimensional. A recent survey (Schultz & Siegel, 1963) suggested that the issues involved in this fact constitute one of the potentially most fruitful areas for further research in criterion development. In this frame of reference, two converging lines of progress led to the work described here.

On the one hand, several recent studies by the authors (Schultz & Siegel, 1961; Siegel & Schultz, 1962) have centered on the utility of the Thurstone and Guttman scaling techniques in the construction of job performance criterion instruments. Since the scaled check lists which were devised provide a practical and efficient method for personnel evaluation, it seemed reasonable to examine the possible contributions of other aspects of scaling methodology, particularly with reference to the important area of criterion dimensionality.

The second emerging research development which suggested the present work was the growing body of literature concerning the theory and application of multidimensional scaling analysis. Originally developed by Richardson (1938), this expansion of

basic psychophysical scaling has recently been studied and extended in some detail by a number of research workers.

The purposes of the present study were to: (a) explore the feasibility of applying standard multidimensional scaling procedures to a job task constellation and (b) determine the number and the nature of the dimensions of a specific Naval job.

MULTIDIMENSIONAL SCALING ANALYSIS

The two central problems in multidimensional scaling analysis are the determination of: (a) the minimum dimensionality of a given set of stimuli and (b) the scale value of each stimulus on each of the dimensions. The specific experimental and computational procedures used have been described in detail by Torgerson (1952, 1958), Messick (1956a, 1956b), and others.

As Gulliksen (1961) has pointed out, the basic judgment upon which the whole structure of multidimensional scaling analysis rests is a very simple one. In order to obtain estimates of the "psychological distances" among the various stimuli in a set, most experimenters (*Es*) have asked the subjects (*Ss*) (judges) to indicate in some manner the degree of overall similarity between each stimulus pair. The methods for obtaining and scaling these distance judgments are generally analogous to the classical psychophysical scaling techniques.

If the obtained scaled values can be taken as measures of the interstimulus distances

¹ This study was performed under Contract Nonr 2279(00) between Applied Psychological Services and the Personnel and Training Branch, Psychological Sciences Division, Office of Naval Research. The authors express their indebtedness to R. Trumbull, G. Bryan, J. Nagay, G. D. Mayo, P. Federman, and D. R. Saunders for their advice and assistance.

in a Euclidean space, the analytical problem then becomes the determination of the number of axes in that space and the projections of the stimuli on these axes. In these final stages, multidimensional scaling analysis uses factor analysis methods. As in factor analysis, for example, the pattern of scale values (loadings) of the stimuli on each dimension presumably enables *E* to attach meaning to, and so to name, the dimensions.

The techniques have been applied to a wide variety of problems. The early work on colors by Richardson (1938) and on relations between nations by Klingberg (1941) has been followed more recently by applications to such areas as attitudes (Abelson, 1954; Messick, 1954, 1956a), personality (Jackson, Messick, & Solley, 1957), jobs (Reeb, 1959), and facial expressions (Abelson & Sermat, 1962), among others, in addition to further work in color (Helm, 1959; Messick, 1956c; Torgerson, 1951, 1952).

Multidimensional scaling differs from unidimensional scaling in one very significant respect. In the typical unidimensional experiment, the scales or dimensions are presented to judges who are asked to order the stimuli on the dimensions as defined by *E*. In multidimensional scaling no such a priori assumptions or definitions are made. Rather, the purpose of the analysis is to discover the number and characteristics of the underlying dimensions which may be justified by the empirical data. Multidimensional scaling analysis also differs from certain other techniques like factor analysis in that the results of multidimensional scaling analysis grow out of the *perceptions* of the *Ss* who make the similarity judgments. The organization of the field that it produces is, therefore, the structure *as perceived by those judges*.

RELEVANCE OF MULTIDIMENSIONAL SCALING METHODS FOR JOB PERFORMANCE MEASUREMENT

In areas where the variables are complex and the dimensions unknown or doubtful, it would seem particularly appropriate to delineate the variables through multidimensional scaling analysis rather than to establish the dimensions arbitrarily. The research in areas of fairly well established dimension-

ality, particularly color, has been cited as evidence of the validity of the multidimensional methods. In view of this validity, Messick (1956c) has suggested that "it would now seem reasonable to apply these procedures for purposes of exploration and discovery in areas of unknown dimensionality [p. 374]."

Thus, multidimensional scaling analysis would appear to be particularly relevant for the development of job performance criteria. The fact that the results of multidimensional scaling reflect the perceptions of the judges introduces an important element into criterion research. For example, in developing measures of job performance, the perceptions of supervisory personnel would seem to merit paramount attention. Yet, for a criterion to be acceptable and meaningful to everyone concerned, there should probably be no serious conflict between the view of the job held by the supervisors and the view held by subordinates.

Delineation of the basic dimensions of a job would serve a number of important purposes. Once the characteristics of the job performance dimensions are known, it should be possible to build unidimensional performance criterion instruments, scaled according to the Thurstone and/or Guttman requirements, by methods previously developed (Schultz & Siegel, 1961; Siegel & Schultz, 1962). Unidimensional, scaled instruments of this type would be useful for the evaluation of the job performance of individuals. In addition, knowledge of the job dimensionality would provide a sound basis for performance test construction, for training device and training-program development, for job analysis, and for work in other related problem areas.

MULTIDIMENSIONAL SCALING FORM

For the present research, interest was centered on the job performance structure of striker and petty officer, third class, Naval aviation electronics technicians. These are men at the entry levels of this job specialty. Therefore, a list of behaviorally oriented job tasks was desired which would be inclusive of all the kinds of work performed by the men in this job but which would not be so detailed as to require an impossibly large

number of similarity comparisons or as to make the judgmental process unreasonably cumbersome.

Three previous studies (Richlin, Siegel, & Schultz, 1960; Schultz & Siegel, 1961; Siegel & Schultz, 1962) had delineated the tasks performed by technicians at this level. Starting from this base of experience, a tentative list was prepared in accordance with the requirements described above. The tasks were stated in terms of operations or "processes" without reference to specific equipment, for two reasons: (a) to avoid the generation of either general or specific "equipment" factors, and (b) so that all the work done by the four types of specialists in this job could be included.

The preliminary list of 40 tasks was reviewed by 23 instructors in the aviation electronics technician school at the Naval Air Technical Training Command. They were asked to indicate the specific tasks done by a representative aviation electronics technician in the Fleet at the entry job levels (striker and petty officer, third class). Space was provided for the addition of any job tasks not included in the preliminary list. These men, all recently arrived from Fleet duty, were chief petty officers or petty officers, first or second class. Their qualifications for responding to such a request included an average of about $7\frac{1}{2}$ years of military experience in electronics or electrical work and about $5\frac{1}{2}$ years of experience as an aviation electronics technician. In addition, during their careers they had been assigned as aviation electronics technicians to an average of about two and one-quarter different squadrons.

In his introductory remarks, the administrator stressed that a picture of the job as it *is* done was wanted, rather than as it is *supposed* to be done. After completion of the form, an informal discussion was held with the groups to determine their estimate of the overall completeness of the list and to obtain comments about such matters as the wording of the tasks. The consensus seemed to be that their selections accurately and adequately reflected the work done by aviation electronics technicians; only a few minor wording suggestions were made.

No meaningful additional tasks were added by any of these judges in the space for "other" tasks. There was almost unanimous agreement that 18 of the tasks were performed in the Fleet; 19 or more of the judges checked these items. The frequency with which the remaining tasks were checked varied greatly. It seemed reasonable to conclude that these widely experienced men felt that the 18 tasks included all the important work normally and customarily performed by striker and petty officer, third class, aviation electronics technicians in the Fleet. *These 18 tasks, therefore, formed the basis for the multidimensional scaling analysis:*

1. Performing variety of "housekeeping" duties such as cleaning shop, repairing tools, and so forth
2. Performing routine line operations
3. Standing watch
4. Performing minor inspections of avionic equipments
5. Performing intermediate inspections of avionic equipments
6. Performing preflight inspections of avionic equipments
7. Performing postflight inspections of avionic equipments
8. Operating avionic equipments
9. Using safety precautions on equipment
10. Using proper safety precautions for self
11. Removing malfunctioning parts/equipment from planes
12. Replacing repaired parts/equipment in planes
13. Performing preventative maintenance on avionic equipments
14. Following block diagrams for avionic equipments
15. Using schematics for standard circuits in avionic equipments
16. Making out reports (failure, and so forth)
17. Using inspection and operation manuals
18. Operating standard test equipment for determining malfunctions in avionic equipments

The stimulus material was arranged in booklet form. At the top of each page in the booklet, one of the 18 tasks was shown. Below it at the left side of the page, the remaining 17 tasks were listed in a random order which was varied from one page to another. A scale running from 1 to 11 appeared to the right of each of the 17 tasks. The Scale Points 1 and 2 were described at the top of the page as representing a judgment of "very similar"; Points 3, 4, and 5 as representing "moderately similar"; Points

7, 8, and 9 as representing "moderately different"; and Points 10 and 11 as representing "very different." Scale Point 6, in the middle of the range, was not described in verbal terms. The directions asked *S* (judge) to compare each task listed with the one shown at the top of the page and then to "indicate by a check in the appropriate column to the right how *similar or different* the two tasks are."

In addition to the random order of tasks on each page, the order of pages in the booklet was determined from a table of random numbers. Four different random page orders were used, the forms being intermixed for administration to *Ss*.

SUBJECTS AND ADMINISTRATION

The *Ss* for this study were 31 chief petty officers and 34 petty officers, first class, in the Naval aviation electronics technician job specialty, assigned to 14 different squadrons at two locations. Their military experience in electronics or electrical work averaged 11.2 years. They had been aviation electronics technicians for an average of 8.3 years and had been assigned as aviation electronics technicians to an average of 3.8 squadrons.

The multidimensional scaling forms were administered to *Ss* in several groups, the largest of which numbered about 20. The booklets were essentially self-administering. There were no time limits but almost all *Ss* had finished in about an hour.

The *Ss* were able to understand their task easily. Most of them proceeded without difficulty and there were almost no questions. In informal conversation with some of the *Ss* as they left the session, several said they felt that they were able to carry out the judging assignment well, although they did not see what ultimate purpose the data would serve.

MULTIDIMENSIONAL SCALING ANALYSIS

The method of equal appearing intervals was used to scale the obtained interstimulus distance judgments. The scale value for each pair of job tasks was taken as the median of the values checked along the similarity scale by *Ss*. Since each pair was judged twice

—once with the first member (A) at the top of the page and once with the second member (B) at the top of the page—both the A-to-B and B-to-A distance judgments were obtained.

The two types of judgments were compared in order to answer the question of whether or not the data met the requirements of "point" items. For point items, the psychological distance to B as judged from A should be the same as the psychological distance to A as judged from B. This non-directional distance characteristic is required if the data are to be considered in terms of a Euclidean space model. The reliability of the judgments is also involved in this comparison, of course. To the extent that the A-to-B and B-to-A scale values are similar, it may be said that the judgments are reliable and meet one of the model's basic specifications. Using attitudinal content, Messick (1956a) found satisfactory agreement between corresponding A-B and B-A distance judgments.

The mean of the 153 A-B interstimulus distance estimates was 5.51 and the standard deviation was 2.22. For the 153 B-A estimates, the mean was 5.49 and the standard deviation was 2.23. The Pearson product-

TABLE 1
FINAL MATRIX OF PROJECTIONS OF TASKS
ON DIMENSIONS

Task number	Dimension number			
	1	2	3	4
1	4.06	-2.61	-0.76	0.87
2	2.12	0.67	0.41	0.27
3	2.69	-4.07	-4.42	1.51
4	0.17	1.26	0.10	0.07
5	-0.44	1.37	0.73	-0.64
6	0.47	1.81	0.09	0.35
7	0.60	1.97	0.19	0.08
8	-1.10	2.71	-0.43	0.72
9	-0.07	0.11	-0.12	1.92
10	0.51	0.07	-0.19	2.37
11	0.71	-0.46	3.14	0.40
12	0.69	-0.03	2.70	0.21
13	-0.67	0.34	1.25	-0.11
14	-3.19	-1.48	0.33	-1.57
15	-3.85	-0.88	0.22	-0.95
16	1.76	-2.33	-2.42	-4.35
17	-1.59	0.79	-0.62	-2.34
18	-2.86	0.76	-0.18	1.19

moment correlation between the scale values for the same distance estimated from opposite directions was calculated to be $+0.92$. Thus, it seems evident that the judgments were made reliably and that the estimates of the psychological distances between tasks were not affected by the direction in which they were made. Therefore, the A-to-B and B-to-A values for each task pair were averaged and the average was taken as the relative intertask distance for use in the multidimensional scaling analysis.

Since the basic theorems underlying multidimensional scaling (Young & Householder,

TABLE 2

TASKS WITH HIGHEST PROJECTIONS ON DIMENSION 1

Task number	Loading	Task
1	+4.06	Performing variety of "housekeeping" duties such as cleaning shop, repairing tools, etc.
3	+2.69	Standing watch
2	+2.12	Performing routine line operations
16	+1.76	Making out reports (failure, etc.)
17	-1.59	Using inspection and operation manuals
18	-2.86	Operating standard test equipment for determining malfunctions in avionic equipments
14	-3.19	Following block diagrams for avionic equipments
15	-3.85	Using schematics for standard circuits in avionic equipments

1938) assume absolute rather than relative interstimulus distances, it was necessary to determine the constant that should be added to the obtained scale values in order to convert them from an interval scale to a ratio scale. The general solution to the additive constant problem proposed by Messick and Abelson (1956) was applied to the data. This produced a negative constant of -1.02 . Since the smallest judged distance was $+1.06$, the corrected smallest intertask distance became almost zero.

The usual multidimensional scaling analytical techniques were then applied to the corrected distance matrix. The first step was to convert this intertask distance matrix to

TABLE 3

TASKS WITH HIGHEST PROJECTIONS ON DIMENSION 2

Task number	Loading	Task
8	+2.71	Operating avionic equipments
7	+1.97	Performing postflight inspections of avionic equipments
6	+1.81	Performing preflight inspections of avionic equipments
16	-2.33	Making out reports (failure, etc.)
1	-2.61	Performing variety of "housekeeping" duties such as cleaning shop, repairing tools, etc.
3	-4.07	Standing watch

a matrix of scalar products of the vectors to points with an origin at the centroid of all the stimuli. This matrix of scalar products was then factored by the method of principal components. The rank of the resulting factor matrix was four. The four axes of the dimensional structure were rotated to orthogonal, simple structure according to the normal equamax criterion, an analytical solution to the rotation problem described by Saunders (1962). The final matrix of projections of the stimuli (tasks) on the rotated axes is presented in Table 1.

INTERPRETATION OF DIMENSIONS

The tasks with the highest projections (loadings) on Dimension 1 are given in Table 2. The tasks with negative loadings seem to involve higher-level understanding of the principles of avionic circuitry, while those with positive loadings are routine, intellectually nondemanding activities. This

TABLE 4

TASKS WITH HIGHEST PROJECTIONS ON DIMENSION 3

Task number	Loading	Task
11	+3.14	Removing malfunctioning parts/equipment from planes
12	+2.70	Replacing repaired parts/equipment in planes
16	-2.42	Making out reports (failure, etc.)
3	-4.42	Standing watch

TABLE 5

TASKS WITH HIGHEST PROJECTIONS ON DIMENSION 4

Task number	Loading	Task
10	+2.37	Using proper safety precautions for self
9	+1.92	Using safety precautions on equipment
14	-1.57	Following block diagrams for avionic equipments
17	-2.34	Using inspection and operation manuals
16	-4.35	Making out reports (failure, etc.)

dimension has, therefore, been tentatively labeled "electro-comprehension."

Table 3 contains the tasks with the highest loadings on the second dimension. Since the three positive loadings grow out of simple operation and inspection tasks, this dimension has been called "equipment operation and inspection (routine)."

As can be seen from the tasks listed in Table 4, the positive direction of Dimension 3 relates to the removal and replacement aspects of repairing equipment in planes. It is, therefore, designated "electro-repair (simple)."

The tasks for which the loadings are highest on the fourth dimension are included in Table 5. The two positive loadings are for tasks which involve the use of safety measures, relating to both the worker and the equipment, as opposed to other aspects of the job. The name selected for this dimension is "electro-safety."

DISCUSSION AND CONCLUSIONS

It should be mentioned, in considering the adequacy of the dimensional structure described here, that the findings of a multi-dimensional scaling analysis, like those of factor analysis, are limited by the input data. In the present study considerable effort was expended on developing a task list that included all the important types of work performed by striker and petty officer, third class, aviation electronics technicians. A more detailed or a less detailed list or one oriented in another manner, such as around the

specific equipment used in the job, might have been employed. The structure produced here directly reflects the task list as developed, at that level of abstraction. Analysis involving other aspects of the job, such as the equipment used, would provide other types of dimensions which might be of value when viewed in conjunction with these task dimensions.

The four dimensions describe activities which are reasonable and meaningful. The operation, inspection, and repair functions are obviously important parts of the aviation electronics technician's job. Strikers and third class petty officers would be working at the lower levels of complexity in these areas, although safety practices and some understanding of the principles of electronic circuitry would be required. In view of these results, it seems reasonable to conclude that it is feasible and fruitful to apply multi-dimensional scaling techniques to the analysis of job performance.

Furthermore, the four dimensions extracted appear to be amenable to unidimensional scaling, i.e., they possess characteristics which would seem to make it possible to develop unidimensional scaled criterion instruments to measure each of them. Since the dimensions extracted in the present study represent the underlying structure of the job performed by aviation electronics technicians, as perceived by their supervisors, the next logical step in measuring that job performance would appear to be the construction of unidimensional scales on each of the dimensions. As mentioned earlier, methods for developing job task performance criterion instruments which meet the Thurstone and/or Guttman scalability requirements are available from previous studies (Schultz & Siegel, 1961; Siegel & Schultz, 1962). If unidimensional, scaled instruments are constructed, it will then be possible to evaluate the job performance of individuals on each of the orthogonal dimensions seen by supervisors as constituting the job.

REFERENCES

ABELSON, R. P. A technique and a model for multi-dimensional attitude scaling. *Publ. Opin. Quart.*, 1954, 18, 405-418.

- ABELSON, R. P., & SERMAT, V. Multidimensional scaling of facial expressions. *J. exp. Psychol.*, 1962, **63**, 546-554.
- GULLIKSEN, H. Linear and multidimensional scaling. *Psychometrika*, 1961, **26**, 9-25.
- HELM, C. E. *A multidimensional ratio scaling analysis of color relations*. Princeton: Educational Testing Service, 1959.
- JACKSON, D. N., MESSICK, S. J., & SOLLEY, C. M. A multidimensional scaling approach to the perception of personality. *J. Psychol.*, 1957, **44**, 311-318.
- KLINGBERG, F. L. Studies in measurement of the relations between sovereign states. *Psychometrika*, 1941, **6**, 335-352.
- MESSICK, S. J. The perception of attitude relationships: A multidimensional scaling approach to the structuring of social attitudes. Unpublished doctoral dissertation, Princeton University, 1954.
- MESSICK, S. J. The perception of social attitudes. *J. abnorm. soc. Psychol.*, 1956, **52**, 57-66. (a)
- MESSICK, S. J. Some recent theoretical developments in multidimensional scaling. *Educ. psychol. Measmt.*, 1956, **16**, 82-100. (b)
- MESSICK, S. J. An empirical evaluation of multidimensional successive intervals. *Psychometrika*, 1956, **21**, 367-375. (c)
- MESSICK, S. J., & ABELSON, R. P. The additive constant problem in multidimensional scaling. *Psychometrika*, 1956, **21**, 1-15.
- REEB, M. How people see jobs: A multidimensional analysis. *Occup. Psychol.*, 1959, **33**, 1-17.
- RICHARDSON, M. W. Multidimensional psychophysics. *Psychol. Bull.*, 1938, **35**, 659-660.
- RICHLIN, M., SIEGEL, A. I., & SCHULTZ, D. G. *Post-training performance criterion development and application: Development and application of a TBCL criterion to the SESR program for aviation electronics technicians*. Wayne, Pa.: Applied Psychological Services, 1960.
- SAUNDERS, D. R. Trans-varimax: Some properties of the ratiomax and equamax criteria for blind orthogonal rotation. *Amer. Psychologist*, 1962, **17**, 395-396. (Abstract)
- SCHULTZ, D. G., & SIEGEL, A. I. Generalized Thurstone and Guttman scales for measuring technical skills in job performance. *J. appl. Psychol.*, 1961, **45**, 137-142.
- SCHULTZ, D. G., & SIEGEL, A. I. Progress and problems in the measurement of individual differences in on-the-job performance. *Acta psychol.*, Amsterdam, 1963, **21**, 120-156.
- SIEGEL, A. I., & SCHULTZ, D. G. Thurstone and Guttman scaling of job related technical skills. *Psychol. Rep.*, 1962, **10**, 855-861.
- TORGERSON, W. S. A theoretical and empirical investigation of multidimensional scaling. Unpublished doctoral dissertation, Princeton University, 1951.
- TORGERSON, W. S. Multidimensional scaling: I. Theory and method. *Psychometrika*, 1952, **17**, 401-419.
- TORGERSON, W. S. *Theory and methods of scaling*. New York: Wiley, 1958.
- YOUNG, G., & HOUSEHOLDER, A. S. Discussion of a set of points in terms of their mutual distances. *Psychometrika*, 1938, **3**, 19-22.

(Received September 30, 1963)

USE OF A LOGICALLY RELATED PREDICTOR IN DETERMINING INTRAGROUP DIFFERENTIAL PREDICTABILITY

JOHN H. STEINEMANN

*United States Naval Personnel Research Activity, San Diego*¹

The overall validity of a career-intention question for predicting Navy reenlistment was reanalyzed for subgroups selected by another logically related test serving as a measure of predictability. On the assumption that career-intention responses of better informed recruits would be relatively more valid, 21 samples, comprising 13,448 enlisted men, were each trichotomized into High, Middle, and Low subgroups on Naval Knowledge Test (NKT) scores. The validity of the career question for the High group was equal to, or larger than the validity for the total group in 19 of the 21 samples. The results generally confirmed that test validity for total groups may be improved for subgroups identified as more predictable by another relevant measure.

Ghiselli (1956, 1960) and others have shown that even though the validity coefficient of a test for a given group may be negligible, there is a possibility that for certain individuals, or subgroups of individuals, prediction of the criterion may be made more accurately than is indicated by the overall predictor validity. The approach involves the use of another test as a "predictability" measure to select individuals whose predictability, in terms of the correspondence of their standard scores on predictor and criterion, is higher than that of the group as a whole.

Ongoing research in the prediction of first-term reenlistment in the United States Navy provided the opportunity to investigate the possibility of differential intragroup predictability, using a similar approach, for a relatively large number of cases over 21 different samples. Unlike many of the previously reported investigations the criterion in this case was dichotomous, i.e., reenlistment or nonreenlistment in the Navy. Of 12 individual measures which had been obtained for all cases during recruit training, the two variables most logically related to the criterion were the response made to a Biographical Informa-

tion Blank (BIB) question² concerning the recruit's career intention and the score on a 45-item Naval Knowledge Test (NKT). It was these two predictors and their interrelationship which were the primary concern of this report.

METHOD

The total sample comprised 13,448 Navy enlisted men for whom predictor measures had been obtained during recruit training, and for whom criterion information (i.e., reenlistment decision made approximately 4 years later) was also available. The total sample included 21 different enlisted ratings, each of which required training at a Navy Class "A" school.

The validity of the BIB item and the NKT for predicting reenlistment decisions was determined by computing the point-biserial correlation of each of these two predictors with the dichotomized criterion. The NKT variable was the total raw score for the 45 test items. The BIB item response was coded on a five-point scale, representing a continuum from negative, through undecided, to positive career intention as follows: (response A = 1, B = 2, E = 3,

² BIB item number 7

"Which statement best describes your plans for staying in the Navy?"

- A. I am certain I will leave the Navy at the end of this enlistment.
- B. I will probably leave the Navy at the end of this enlistment.
- C. I will probably stay in the Navy for at least another enlistment.
- D. I plan to make a career of the Navy.
- E. I haven't made up my mind yet, or I don't know.

¹ The opinions and conclusions expressed herein do not necessarily reflect the opinions of the Chief of Naval Personnel or the Department of the Navy.

TABLE 1

CORRELATIONS BETWEEN BIB CAREER ITEM AND REENLISTMENT DECISION FOR
RATING SAMPLES TRICHOTOMIZED ON NAVAL KNOWLEDGE TEST SCORES

Sample	Group	r_{pb}	Sample	Group	r_{pb}
Yeoman ($N = 714$)	H	14	Storekeeper ($N = 561$)	H	20
	M	13		M	09
	L	01		L	05
	Total	11		Total	11
Aviation Structural Mechanic (structural) ($N = 592$)	H	19	Air Controlman (tower) ($N = 206$)	H	30
	M	05		M	09
	L	06		L	02
	Total	09		Total	13
Machinist Mate ($N = 816$)	H	16	Hospitalman ($N = 1387$)	H	22
	M	14		M	05
	L	10		L	10
	Total	14		Total	12
Aviation Structural Mechanic (hydraulic) ($N = 867$)	H	23	Electronics Technician (com- munication) ($N = 758$)	H	26
	M	08		M	06
	L	06		L	11
	Total	08		Total	17
Interior Communications Electrician ($N = 577$)	H	16	Air Controlman ($N = 261$)	H	23
	M	-02		M	13
	L	14		L	32
	Total	11		Total	16
Engineman ($N = 275$)	H	15	Aviation Storekeeper ($N = 219$)	H	23
	M	12		M	29
	L	29		L	16
	Total	20		Total	22
Fire Control Technician ($N = 1571$)	H	20	Aviation Machinists Mate ($N = 527$)	H	14
	M	10		M	08
	L	13		L	19
	Total	15		Total	14
Electronics Technician (radar) ($N = 991$)	H	11	Radarman ($N = 578$)	H	16
	M	15		M	-08
	L	02		L	24
	Total	11		Total	09
Machinist Repairman ($N = 367$)	H	31	Aviation Electronics Techni- cian (navigation) ($N = 1052$)	H	15
	M	01		M	16
	L	06		L	10
	Total	06		Total	14
Aviation Electrician's Mate ($N = 613$)	H	26	Dental Technician ($N = 140$)	H	-29
	M	17		M	15
	L	07		L	-09
	Total	19		Total	-08
Aviation Electronics Techni- cian (radar) ($N = 366$)	H	24			
	M	17			
	L	12			
	Total	18			

C = 4, D = 5). The reenlistment criterion was coded 1 for reenlistment and 0 for nonreenlistment.

Correlations obtained for 46 groups (some of the 21 rating samples had been divided into two or more subsamples because of considerations of size or location) indicated a low but fairly consistent validity for the BIB item, (range r_{pb} $-.07$ to $.27$, median $.14$) and an even lower NKT validity, (range r_{pb} $-.07$ to $.22$, median $.08$).

Although the magnitude of the BIB validities was small, the size and number of samples upon which they were obtained suggested that they represented a reliable relationship between the BIB response made as a recruit and eventual reenlistment decision made 4 years later. It may be noted that for practical purposes, any valid career predictors which could be obtained early in recruit training, prior to assignment and training, would be of greatest value in terms of manpower and financial economies. However, the 4 years intervening between recruit training and actual career decisions undoubtedly tends to reduce the validity of any predictor measures. For the BIB item, in particular, it seemed probable that many recruits have little or no prior knowledge of, or experience with, actual Navy life and consequently have little basis on which to make a meaningful response to a question regarding their eventual reenlistment plans. Partial confirmation of this assumption was found in analysis of BIB responses by groups which showed that among recruits responding negatively to the item (i.e., intended to leave the Navy) 13% actually reenlisted, compared with 19.5% for the total group.

It was in order to investigate the possibility that BIB responses made by recruits with relatively more knowledge about the Navy might be more valid than responses made by less informed individuals, that use of the NKT as a predictor of predictability was decided upon. It was hoped that NKT, although having only low validity for the criterion, might be useful in identifying better-informed, hence more predictable, individuals with respect to the criterion.

Accordingly, each of the 21 rating samples was trichotomized on the basis of NKT score into High (H), Middle (M), and Low (L) subgroups of approximately equal size. Point-biserial correlations were then computed between BIB response and the criterion for each subgroup and total group for all 21 samples. It was hypothesized that the BIB-criterion correlations for the H groups, representing recruits best informed about the Navy, would be higher than the BIB validities for the total groups, which included all levels of naval knowledge.

RESULTS AND DISCUSSION

Table 1 presents the results of this analysis in terms of point-biserial correlations of the BIB item with the reenlistment criterion for H, M, and L groups and for the total groups. It may be seen that the BIB-criterion correlations were larger for the H group than for the total group in 17 of the 21 samples, and were equal to the total group in two samples. The BIB validity of the H group was smaller than for the total group in two samples, one of which was the Dental Technician sample of relatively small size in which the relationship was negative.

A summary comparison of the BIB validities for H, M, L and total groups is presented below in terms of the median and range of correlations obtained for all 21 samples.

Group	Range r_{pb}	Median r_{pb}
High	$-.29$ to $.31$	$.20$
Middle	$-.08$ to $.29$	$.10$
Low	$-.09$ to $.32$	$.10$
Total	$-.08$ to $.22$	$.13$

These results generally confirmed the hypothesis that the overall effectiveness of a predictor variable, as indicated by its validity coefficient for total groups, can be improved for subgroups identified as more "predictable" by some logically related measure. In this investigation, prediction of career decisions on the basis of response to a BIB career question was better for subgroups scoring high on NKT than for total groups including all levels of naval knowledge.

REFERENCES

GHISELLI, E. E. Differences of individuals in terms of their predictability. *J. appl. Psychol.*, 1956, 40, 374-380.
GHISELLI, E. E. The prediction of predictability. *Educ. psychol. Measmt.*, 1960, 20, 3-8.

(Received September 5, 1963)

DETROIT, MICHIGAN
PLEASE DO NOT REMOVE

Journal of Applied Psychology

KENNETH E. CLARK, Editor
University of Rochester

Table of Contents

The Influence of Task Complexity and Practice on Performance after Loss of Sleep.....	D. W. J. Corcoran	339
Check-Reading Accuracy as a Function of Pointer Alignment, Patterning, and Viewing Angle.....	Sidney G. Dashevsky	344
Combining Check-Reading Accuracy and Quantitative Information in a Space-Saving Display.....	Sidney G. Dashevsky	348
Validation of a Multiple-Assessment Procedure for Managerial Personnel.....	Paul A. Albrecht, Edward M. Glaser, and John Marks	351
Sensory-Feedback Analysis of Behavior in Stereotelevised Visual Fields.....	Karl U. Smith and John D. Gould	361
Stereoscopic Television Pursuit Tracking.....	John D. Gould	369
Validation of the Minnesota Vocational Interest Inventory for Vocational High School Boys.....	W. Leslie Barnette, Jr., and John N. McCall	378
The Effects of Task and Method of Stimulus Presentation of the Detection of Deception....	Lawrence A. Gustafson and Martin T. Orne	383
Job Characteristics as Satisfiers and Dissatisfiers.....	Frank Friedlander	388
Teacher Accuracy in Assessing Cognitive Visual Feedback from Students.....	Jon Jecker, Nathan Maccoby, Henry S. Breitrose, and Ernest D. Rose	393
Studies in the Reliability and Validity of the Critical Incident Technique.....	Bengt-Erik Andersson and Stig-Göran Nilsson	398

This is the last issue of Volume 48.
Volume Title Page and Contents appear herein.

American Psychological Association

Consulting Editors

GEORGE E. BRIGGS, *Ohio State University*
MARVIN D. DUNNETTE, *University of Minnesota*
NORMAN FREDERIKSEN, *Educational Testing Service*
LEONARD D. GOODSTEIN, *University of Cincinnati*
EDWIN R. HENRY, *Standard Oil Company of New Jersey*
JOHN HOLLAND, *American College Testing Program*

CLIFFORD E. JURGENSEN, *Minneapolis Gas Company*
LAURENCE S. MCGAUGHRAN, *University of Houston*
QUINN MCNEMAR, *Stanford University*
HAROLD F. ROTHE, *Beloit Corporation*
THOMAS A. RYAN, *Cornell University*
ALEXANDER G. WESMAN, *Psychological Corporation*
CLARK L. WILSON, *Harvard Business School*

This journal gives primary consideration to original investigations in any field of applied psychology except clinical psychology, although a descriptive or theoretical article may be accepted if it represents a special contribution in an applied field. Quantitative investigations of interest or value to psychologists working in the following broad fields will be considered: vocational and educational prognosis, diagnosis, and guidance at the secondary and college level; personnel research in business, industry, and government; engineering psychology; industrial working conditions; research on opinion and morale factors; job analysis and classification research; market and advertising research.

Manuscripts must be accompanied by an abstract of 100–120 words typed on a separate sheet of paper. The abstract should conform to the style of *Psychological Abstracts*. Detailed instructions for preparation of the abstracts appeared in the *American Psychologist* (1961, 16, 833), or they may be obtained from the Editor or from the APA Central Office.

Manuscripts should be addressed to the Editor:

Dr. Kenneth E. Clark
College of Arts and Science
302 Morey Hall
University of Rochester
Rochester, N. Y. 14627

All manuscripts must be submitted in duplicate. Original figures are prepared for publication; duplicate figures may be photographs or pencil-drawn copies.

Manuscripts must conform to the style requirements described in the *Publication Manual of the American Psychological Association*.

The following policies govern reprinting of materials copyrighted in APA journals: (a) to require approval to reprint for tables and figures, and for text only if more than 500 words [total from one article] in length; (b) to continue approval to reprint, contingent on the author's approval, for articles reprinted in whole or in major part; (c) to negotiate where possible the dedication of royalties from commercial publishers to the American Psychological Foundation.

Journal of Applied Psychology

Published bimonthly by the
American Psychological Association
Prince and Lemon Sts., Lancaster, Pa. 17604
and 1200 Seventeenth St. N. W.
Washington, D. C. 20036

\$10.00 per volume

\$2.00 per issue

HELEN ORR
Managing Editor

ELIZABETH S. REED
Advertising Manager

FRANCES L. BREWER
Editorial Assistant

Subscriptions, orders, and business communications should be addressed to the American Psychological Association, 1200 Seventeenth St. N. W., Washington, D. C. 20036. Address changes must reach the subscription office by the tenth of the month to take effect the following month. Undelivered copies resulting from address changes will not be replaced; subscribers should notify the post office that they will guarantee second-class forwarding postage. Other claims for undelivered copies must be made within four months of publication.

Second class postage paid at Lancaster, Pa. and at additional mailing offices.

© 1964 by the American Psychological Association, Inc.

Journal of Applied Psychology

VOL. 48, No. 6

DECEMBER 1964

THE INFLUENCE OF TASK COMPLEXITY AND PRACTICE ON PERFORMANCE AFTER LOSS OF SLEEP

D. W. J. CORCORAN¹

Medical Research Council, Applied Psychology Research Unit, Cambridge, England

An experiment was conducted to assess the effects of degrees of task complexity and practice on performance after loss of sleep. The Ss were automatically presented every 7 secs. for 23 mins., with cards containing 6 symbols. A symbol had to be chosen on the basis of certain rules. Some cards required 1 rule, some 2, some 3, and some 4. Group 1 was practiced after normal sleep and tested after 22 and 46 hrs. without sleep. Group 2 was tested without sleep and without previous practice. Group 3 was practiced and tested after normal sleep. Loss of sleep had a greater effect after practice, but no clear differences emerged between the different levels of task complexity.

After describing a series of experiments on the influence of certain parameters on performance after loss of sleep, Wilkinson (1958) concluded that "in a given task situation, the performance of moderately sleep-deprived subjects will tend more to that of normals as (a) the situation becomes less predictable (more complex) and (b) the penalties for failure to predict correctly become greater." Other workers, e.g., Tyler (1947), Williams, Lubin, and Goodnow (1959) have not considered complexity and Kleitman (1939) is of the opinion that the more complex task is likely to be the more sensitive to loss of sleep.

In the present experiment, an attempt has been made to discover in what way (if any) degree of complexity influences the susceptibility of a task to loss of sleep. "Complexity" may usefully be considered to be of three kinds: (a) stimulus complexity, which is roughly equivalent to the amount of information per stimulus group presented to the subject (S), (b) response complexity, which depends upon the degree of skill required to execute a response and (c) stimulus-

response complexity, which is equivalent to the degree of "compatibility" of stimulus and response. Only stimulus complexity and stimulus-response complexity have been considered in the present experiment.

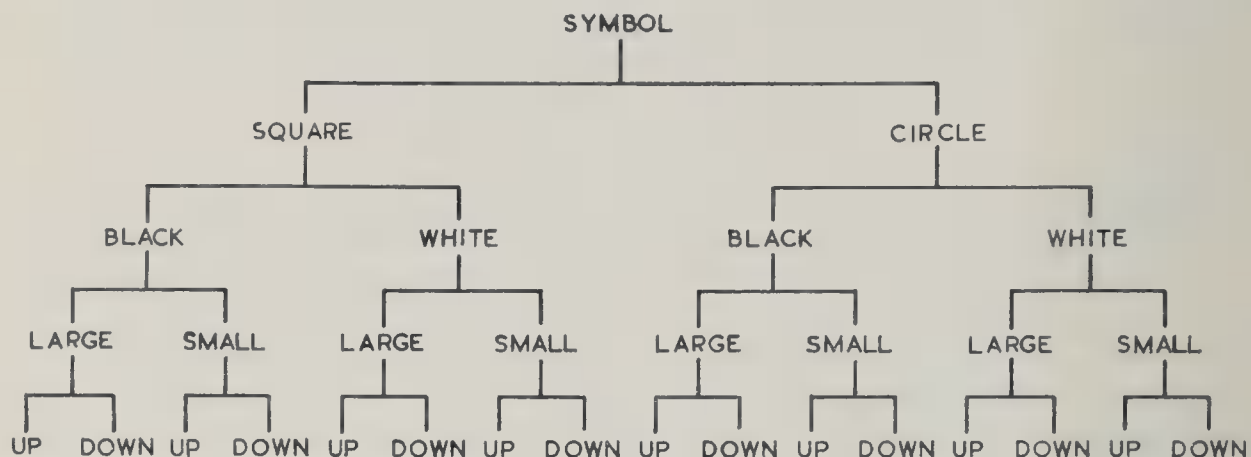
Stimulus complexity was varied directly in four definable steps, whereas stimulus-response complexity was manipulated indirectly by introducing practice as a variable. It was reasoned that whilst the amount of information presented is not influenced by practice, the degree of compatibility of stimulus and response is increased by practice (e.g., Conrad, 1962). Two specific hypotheses were therefore tested: (a) that as the amount of information presented per stimulus group increases, the effect of loss of sleep will become less; (b) that as practice increases, compatibility will increase and therefore the effect of loss of sleep will become greater.

METHOD

Task. Colquhoun's apparatus (Colquhoun, 1961) was used. One hundred cards each $5\frac{1}{2}$ inches \times $\frac{3}{4}$ inch were arranged in sequence around the periphery of a drum. In this way one card at a time was presented to the S through a window for a period of 7 seconds, after which the next card was presented and so on. Six response keys were set horizontally below the window, about 1 inch apart. A card contained six symbols, each of which corresponded to,

¹The author wishes to thank D. E. Broadbent for his advice and the Royal Navy for supplying subjects, equipment, and assistants.

and lay immediately above, one of the keys. The *S* had to select one of the six symbols and record his choice by pressing the appropriate key. The symbols were either squares or circles, black or white, large or small, and located either on an upper or lower position on the cards. Altogether there were 16 possible combinations as outlined below.



The *S* was instructed to choose a single symbol on the basis of the following four rules: (1) choose squares rather than circles, (2) choose black rather than white, (3) choose large rather than small, and (4) choose up positions rather than down positions. Each rule had equal "weight," so that the *S* had to select from among the six symbols the one with the greatest number of positive attributes.

The 100 cards were divided into four sets of 25. Each set involved a different number of rules. The cards in one set required only one rule, as in Figure 1a. This card contains symbols which are

all circles, all small, and all black, but one is higher than the rest; thus according to Rule 4, the higher one has the greatest value. The cards in a second set involved two rules. Thus in Figure 1b all symbols are the same size and in the same position, but differ in color and shape. The third set is exemplified by Figure 1c which required three

rules. The fourth set as in Figure 1d involved four rules. Clearly the amount of information which had to be processed became greater as the number of rules increased.

Two complete versions of the 100 cards were constructed, the second version being exactly equivalent to the first except in the horizontal position of the correct symbol on the card. Within each version this symbol had an almost equal probability of occurring in any horizontal position on the card. The cards were placed in random order on the drum, and this order was changed for each repetition of the test.

Subjects. Eighteen naval ratings between the ages of 18 and 23 served as *Ss*. All had volunteered for work on extended periods without sleep.

Procedure. The *Ss* were divided into three groups of six, as indicated in Table 1. Group 1 carried out the task after normal sleep on two successive afternoons of 1 week, but during the second week was tested on two successive early mornings after approximately 22 and 46 hours of sleeplessness, respectively. Group 2 was tested after loss of sleep during the first week and after normal sleep during the second. Group 3 was tested on two successive afternoons during the first week and this was repeated on the subsequent week. One version of the test was used for each week, with the order of the series changed between the two tests. The drum was allowed two complete revolutions, so that the *Ss* were presented with each card twice, once during the first half of the test and once during the second. The task lasted for about 23 minutes.

The *Ss* were instructed individually on spare cards. They were then tested on four other spare cards, to see if the instructions had been understood.

The present task was one of a number carried

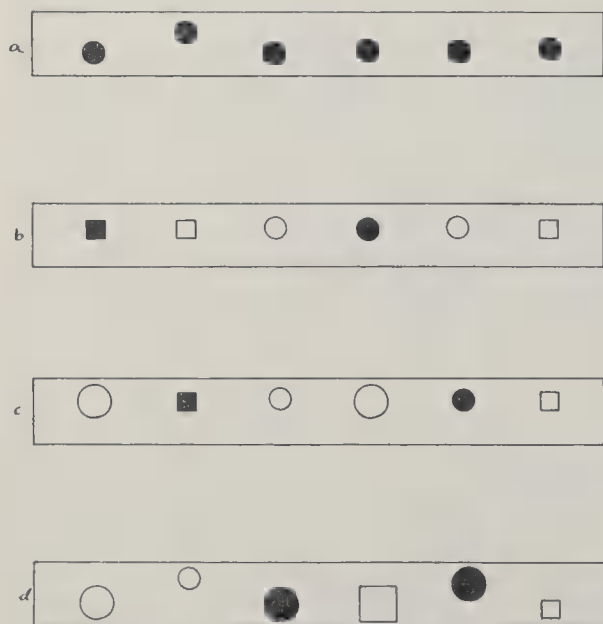


FIG. 1. Examples of the four types of cards used.

TABLE 1
PERCENTAGE ERRORS

Group	Test 1				Test 2				Test 3				Test 4			
	Day 1				Day 2				Night 1				Night 2			
	Rules				Rules				Rules				Rules			
1 <i>M</i> Range and Variance	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
	7.0	13.0	18.0	33.0	2.3	6.7	11.0	24.3	0.7	1.7	8.0	16.9	1.0	8.0	8.7	25.3
	Total	71.0			Total	44.3			Total	27.3			Total	43.0		
2 <i>M</i> Range and Variance	0-42	0-70	4-46	16-70	0-14	0-18	4-32	6-58	0-4	0-2	0-20	0-40	0-4	2-24	0-24	8-64
	73.91				40.27				19.87				36.55			
	Night 1				Night 2				Day 1				Day 2			
3 <i>M</i> Range and Variance	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
	4.0	8.0	14.7	29.7	4.4	6.6	12.4	23.1	4.0	7.7	9.7	25.0	5.0	7.0	9.3	17.7
	Total	56.4			Total	46.5			Total	46.4			Total	39.0		
4 <i>M</i> Range and Variance	0-24	0-40	6-36	12-62	0-24	0-32	0-42	4-56	0-24	0-46	0-40	6-70	0-30	0-38	0-40	4-60
	45.83				49.72				60.74				58.15			
	Day 1				Day 2				Day 3				Day 4			
5 <i>M</i> Range and Variance	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
	0.6	1.0	9.0	24.6	0.3	0.0	4.3	16.3	0.3	3.3	5.0	14.3	0.0	1.6	4.3	7.3
	Total	35.2			Total	20.9			Total	22.9			Total	13.3		
6 <i>M</i> Range and Variance	0-4	0-2	4-18	12-56	0-2	0-0	0-12	2-42	0-2	0-10	0-10	0-36	0-0	0-6	0-16	4-30
	21.47				16.71				16.35				13.88			

out over 60 hours deprivation of sleep. The Ss reported in groups of 6 at the laboratory at 7:00 A.M. on a Tuesday and went without sleep until 7:00 P.M. on the following Thursday. Most of the 60 hours were spent doing laboratory tasks of various kinds, the intervals being filled with games, reading, or having meals. Constant supervision was enforced over the entire period and Ss were aroused as gently as possible when they fell asleep. They were not, however, stimulated whilst the task was in progress. The motivation of the groups was fairly high, since the Ss who showed overt signs of fatigue were generally teased by the more alert members of the group. The experimenter (*E*) and his assistant, however, kept as remote as possible in order that their attitudes to one group should not differ from those to another.

Control tests were conducted in the same order as the experimental, so that the experimental test at say 4 A.M. had its control at 4 P.M. Tests conducted during the first night are therefore influenced by diurnal rhythm and those during the second night by diurnal rhythm and considerable loss of sleep.

RESULTS

Overall Error Score

Table 1 shows that Group 1 improved up to and including the third test, despite the fact that the third test took place at between 2 A.M. and 4 A.M. after about 22 hours of wakefulness. Thus efficiency at this task was not impaired a few hours after normal bedtime, when diurnal rhythm is least favorable. Twenty-four hours later, however, performance had apparently deteriorated; five out of the six Ss in Group 1 showed more errors on the fourth test than on the third. The Mann-Whitney *U* test (in Siegel, 1956) showed that the increase in errors between the third and the fourth test was significantly greater in Group 1 than in Group 3 ($U = 5$, $p = .042$).

The effect of loss of sleep in Group 2 was quite different from that in Group 1. Group 2 who performed the first two tests without sleep showed no evidence of a decline in performance after 46 hours without sleep; in fact there was no difference between the change from the first to the second test in Group 2 and the equivalent changes in Groups 1 and 3. The decline in performance from Test 3 to Test 4 in Group 1 was significantly greater than the change from Test 1 to Test 2 in Group 2 ($U = 5$, $p = .042$). Since performance declined after 46 hours of sleepless-

ness in comparison with 22 hours without sleep only in Group 1, it is clear that the two previous practice sessions had a deleterious effect on this group. However, owing to the poor level of performance on Test 1 and Test 2 in Group 1, and the subsequent effect of practice, the scores on Test 3 and Test 4 are not significantly different from the previous two tests.

Error Scores in Relation to Varying Complexity

Table 1 shows that the errors tended to increase as the number of rules necessary for solution increased.

Two levels of stimulus complexity only were responsible for the general rise in errors from Test 3 to Test 4 in Group 1. These were the cards involving two and four rules. In the two-rule cards the Mann-Whitney *U* test gives a $U = 2$, $p = .004$ whilst in the four-rule cards $U = 7$, $p = .047$; the one-rule and three-rule items showed no difference between Test 3 and 4. Allowing for the fact that these comparisons were made on an ad hoc basis, the estimate of the probability that the four-rule items were affected is probably well above .05. The probability that the two-rule cards were associated with a greater number of errors, however, was probably $< .02$ when the same statistical controls were effected.

DISCUSSION

It is evident that practice enhanced the susceptibility of this task to loss of sleep, corroborating the finding of Wilkinson (1961). This may confirm the hypothesis that an increase in stimulus-response compatibility or a reduction in "response-competition," (which could occur as a result of practice) can increase susceptibility to loss of sleep. However, there is the possibility that the "novelty" or "interest" of a situation is reduced by repeated testing and that it was this factor which was responsible. Unfortunately the present experiment cannot discriminate between these two alternative explanations.

The results are, however, important in respect to the finding of Mackworth (1950) who showed that the performance of the

skilled (practiced) worker was the least affected by room temperature. The present experiment has shown that a greater degree of practice is associated with *increased* susceptibility to loss of sleep, indicating that heat and loss of sleep, although commonly classed together as psychological stresses are in fact idiosyncratic in their effects upon performance (Pepler, 1959).

The relationship between the four levels of complexity and loss of sleep is peculiar. If, as hypothesized, the amount of information to be processed reduced the susceptibility of tasks to loss of sleep, then the four-rule items should be least affected by loss of sleep, the one-rule items should be the most affected, whilst the intermediate two- and three-rule items should fall between these extremes. In fact it was found that the two and perhaps the four-rule items were affected while the rest were not. The present experiment has not therefore demonstrated that stimulus complexity influences the magnitude of the effect of loss of sleep on performance.

REFERENCES

- COLQUHOUN, W. P. The effect of unwanted signals on performance in a vigilance task. *Ergonomics*, 1961, **4**, 41-51.
- CONRAD, R. Practice, familiarity and reading rate for words and nonsense syllables. *Quart. J. exp. Psychol.*, 1962, **14**, 71-76.
- KLEITMAN, N. *Sleep and Wakefulness*. Chicago: Univer. Chicago Press, 1939.
- MACKWORTH, N. H. Researches on the measurement of human performance. *Med. Res. Council, Special Rep.*, 1950, Ser. No. 268.
- PEPLER, R. D. Warmth and lack of sleep: Accuracy or activity reduced. *J. comp. physiol. Psychol.*, 1959, **52**, 446-450.
- TYLER, D. B. Effects of amphetamine sulphate and some barbiturates on fatigue produced by prolonged wakefulness. *Amer. J. Physiol.*, 1947, **150**, 253-262.
- WILKINSON, R. T. The effect of lack of sleep on performance. *Med. Res. Council, Appl. Psychol. res. Unit, Great Britain*, 1958, No. 323.
- WILKINSON, R. T. Interaction of lack of sleep with knowledge of results, repeated testing and individual differences. *J. exp. Psychol.*, 1961, **62**, 263-271.
- WILLIAMS, H. L., LUBIN, A., & GOODNOW, J. L. Impaired performance with acute sleep loss. *Psychol. Monogr.* 1959, **73**(14, Whole No. 484).

(Received September 27, 1963)

CHECK-READING ACCURACY AS A FUNCTION OF POINTER ALIGNMENT, PATTERNING, AND VIEWING ANGLE

SIDNEY G. DASHEVSKY ¹

University of Rochester

Prior studies have shown that detecting a single deviant dial in a 4 × 4 matrix can be markedly improved by aligning the null position of the pointers. In the present study, efficiency in a similar task was rendered 85% more efficient by continuing the line formed by the pointers across the entire panel face. The deviant dial then appeared as a break in a line, a finding consonant with the Gestalt principle of figural continuity. No significant difference was found between the 9 and 12 o'clock orientations for null pointers, nor did any significant difference emerge when displays were presented to the front and the sides of Ss.

In applications involving many dials which must be intermittently but accurately monitored in a brief period of time, considerations of safety take precedence over those of expense and habitual design. The demands made on pilots monitoring multiengine aircraft in flight led the Air Force to study the feasibility of modifying dial displays for greater efficiency. The bulk of this work was done between 1948 and 1953.

A number of researchers found that detection of errors in clusters of dials could be made more rapid and reliable if all pointers were aligned during normal operation. Various alignment positions were investigated, among them pointer normal or null orientations at 12, 3, 6, or 9 o'clock. These would have all pointers of a cluster aligned vertically upward, horizontally to the right, vertically downward or horizontally to the left, respectively. The majority of these studies employed a 4 × 4 matrix of 16 dials exposed for 0.50 or 0.75 second. In addition, one author attempted to introduce patterning into his matrices by orienting the pointers toward the center of a 2 × 2 subdisplay (Johnsgard, 1953). In two studies (Warrick & Grether, 1948; White, 1951), the 9 o'clock orientation was found to permit greater error detection than other orientations. This may, however, be due to the fact that the required response in one study

(White, 1951) was more compatible with the 9 o'clock position than any other. The response involved the flipping down of a key if the dial showed "too much" and the flipping up of a key when it read "too little." Generally, either the 9 or 12 o'clock orientation yielded optimum error detection, with the 3 o'clock alignment evincing a marked decrement in performance. In the study attempting to utilize subgrouping, the method used resulted in a doubling of the time required to read the grouped dials. It is the opinion of the writer that Johnsgard had an excellent idea in suggesting that increased organization of the display might reduce errors, but that he chose an inadequate means of grouping. The sole organizing factor employed by Johnsgard was the pointers themselves—i.e., the variable portion of the display.

The present studies were designed to determine whether or not multiple-dial, check-read displays could be arranged in a manner that would still further increase the probability of detecting errors.

The first experiment is a comparison of the 9 and 12 o'clock orientations in order to estab-

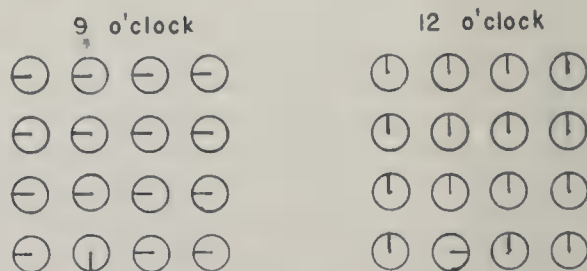


FIG. 1. Dial displays for Experiment 1.

¹ This study was conducted at the Human Engineering Laboratory, Aberdeen Proving Ground, Maryland.

The author wishes to express his appreciation for suggestions made by S. J. Glucksberg, of the above installation.

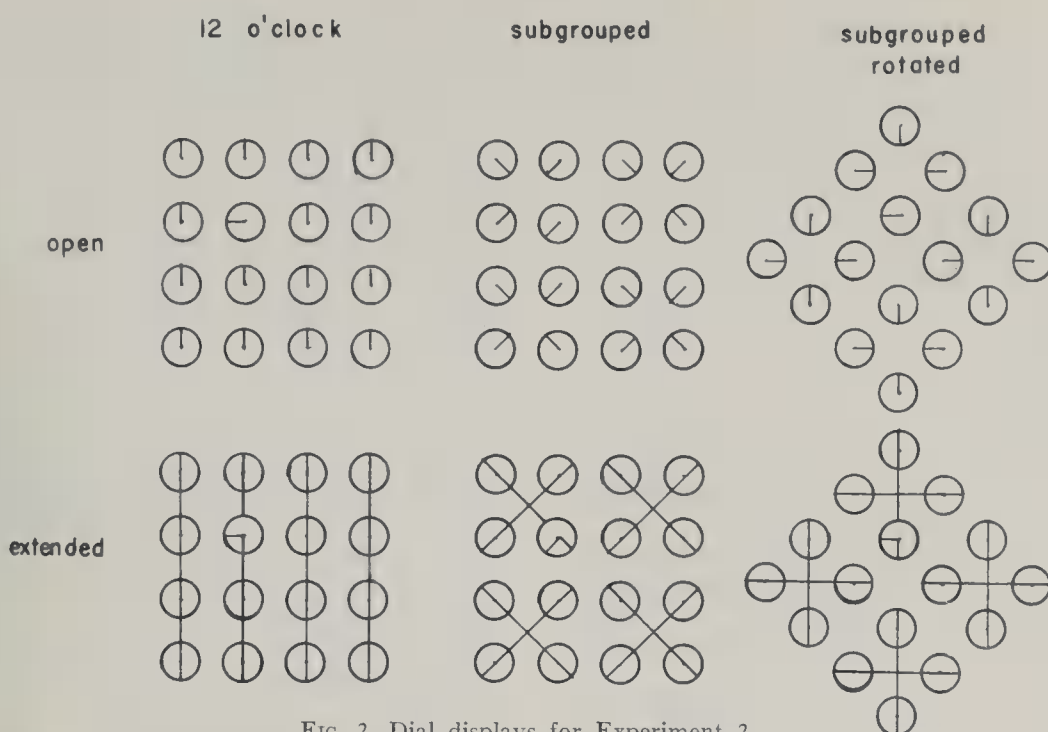


FIG. 2. Dial displays for Experiment 2.

lish an optimum basis for further comparison (see Figure 1). In addition, the effects of presenting displays directly in front of and to the subject's (*S*'s) side are compared. This was done to investigate the possibility that eye movement between display and work surface might be more natural for some pairs of conditions than others. Specifically, it seemed possible that the 9 o'clock orientation would prove superior when the display was at the *S*'s side and the 12 o'clock orientation would prove superior when presented to the front, since eye movements from display to work surface would be most direct under these conditions.

The second experiment was designed to compare the mode of alignment found best with displays grouped: (*a*) according to Johnsgard's method, (*b*) as in (*a*), but rotated 45 degrees to provide a possibly more natural configuration, and (*c*) grouped according to Gestalt theory principles of good form and closure and utilizing a means of organization other than the pointers alone.

EXPERIMENT I

This experiment compared the accuracy of error detection of 9 and 12 o'clock alignment modes, and under frontal or peripheral presentation.

Method

Subjects. A total of 25 male civilian, enlisted and officer personnel of the Human Engineering Laboratory served as *Ss*. All were college students or graduates between 18 and 30 years of age having at least 20/30 vision.

Apparatus. The *Ss* were seated in a large darkened room 8–10 feet from a projection screen at the front. Cards simulating 4 × 4 displays were projected from 12 feet by a Beseler opaque projector, over the lens of which a Wollensak Alphax shutter had been fitted. The *Ss* recorded responses on record forms representing 4 × 4 matrices. The record forms were held in illuminated clipboards.

Procedure. The *Ss* were told that one of the 16 dials shown in any display might be deviant by 90 or 270 degrees and that their task was to indicate on the response sheet whether all pointers were aligned or not. If a display was perceived as nonaligned, the *S* was to indicate the deviant dial.

For side presentation, *Ss* turned their chairs 90 degrees, so that half the *Ss* would find the display on their right, half on their left.

For each of the four conditions (9:00 front, 9:00 side, 12:00 front, 12:00 side), 20 cards were presented sequentially for 0.5 second each. Of any 20 cards, 10 contained one nonaligned dial. These were presented in groups of 10 randomly arranged cards. Particular dials and cards deviant, order of cards within sets and order of presenting sets were all randomized. The *Ss*' attention was directed to the screen by the experimenter's (*E*'s) announcing the number of the card to be flashed 1–2 seconds before presentation.

The *Ss* were run in 3 groups of 8–10. One set

TABLE 1
COMPARISON OF ALIGNMENT AND PRESENTATION MODES

	Raw errors	
	9 o'clock	12 o'clock
Frontal presentation	24.5	21.5
Peripheral presentation	17.5	14.5

of 10 cards was presented to provide familiarity with the task, after which each *S* received all conditions. Running time per group averaged 30 minutes.

Results. Number of errors was used as the dependent variable, error being defined as missing a deviation or reporting one for a null display. In addition, ascribing a deviation to the wrong dial was scored one-half error.

Since *Ss* were used as their own controls, the *t* test for pairs of correlated means (Guilford, 1956, p. 220) was chosen as an appropriate statistic. The comparison between the 9 and 12 o'clock modes of alignment yielded a nonsignificant *t* of 1.23. The comparison between the frontal and peripheral modes of presentation yielded a *t* of 1.82, also nonsignificant. Total raw errors are presented in Table 1.

EXPERIMENT II

The major variable studied was that of continuing across the dial face and panel the lines made by normally aligned pointers. Any deviation would then appear as either a break in, or shortening of, a line. The 12 o'clock arrangement was selected for comparison as it showed a slight superiority in the first experiment. This was compared with an open-x arrangement (2 × 2 subgrouping of dials) identical to that employed by Johnsgard, and with the open-x display rotated 45 degrees to produce an upright cross in each subgroup. Each of these three modes of alignment was then compared with the same mode in which the lines made by pointers were extended across the dial face and panel (see Figure 2).

Method

Subjects. Eighteen *Ss* having the same characteristics as those used in the first experiment served. Some (whose error scores were not significantly different from the others) had served in the first experiment.

Apparatus. The physical arrangements were those used in the first study. In Experiment I, however,

dials simulated white faces on black panels, in this study dials simulated white faces on white panels.

Procedure. The *S's* task remained unchanged: 20 cards per condition being presented for 0.5 second each in two groups of 10 cards. The following changes were made: (a) 15 of each set of 20 cards had one deviant dial, (b) the 180 degrees deviation was used, as well as the 90 and 270, (c) only gross errors were scored (missing deviations or reporting nonexistent ones). Dials deviant, direction of deviation, order of cards within sets, and order of presentation of sets were randomly determined.

The *Ss* were given 15 cards demonstrating all conditions as pretraining.

Results. Total raw errors are presented in Table 2. The *t* test for pairs of correlated means was used, with the following results: (a) extension of pointers reduced errors for all modes of patterning by almost 85%; *t* = 13.00, *p* < .01; (b) the 12 o'clock mode elicited fewer errors than either the subgrouped or subgrouped rotated modes; *t's* = 5.73 and 8.04, respectively, (*p* in both cases < .01); (c) rotating the subgrouped display increased the number of errors made, *t* = 2.40, *p* < .05; see Table 2.

CONCLUSIONS

The following conclusions emerged:

(a) there seems to be no consistent difference between using the 9 or 12 o'clock mode of alignment when a 4 × 4 matrix is used; although in some prior studies the 9 o'clock alignment resulted in superior performance;

(b) there seems to be no interaction of 9 or 12 o'clock alignment and presentation of the display at the *S's* front or side;

(c) extending the lines made by pointers completely across the panel can result in a dramatic reduction of errors, since single deviant pointers then show up as breaks in a line. The simpler the resulting pattern, the more accurate the performance.

It is of interest to note that many *Ss* complained that the shutter was inaccurate—that certain displays seemed to be exposed for

TABLE 2
COMPARISON OF PATTERNING MODES

	Raw errors		
	12 o'clock	Sub-grouped	Sub-grouped rotated
Open	53	193	201
Extended	8	15	41

much shorter periods than others. Upon questioning, it developed that those displays perceived as too briefly exposed were invariably of the open-x type which would, within this framework, be considered poor Gestalten.

Where most researchers have found detection of a 180 degrees deviation most difficult, this deflection was readily noticed when extensions of pointers were introduced. It would be more difficult to detect where it occurred on a peripheral dial, since this deviation would then appear as a shortening of a line, rather than as a break, but by using extension, detection probability is markedly improved.

REFERENCES

- BONEAU, C. A. The effects of violations of assumptions underlying the *t* test. *Psychol. Bull.*, 1960, 57, 49-64.
- FITTS, P., & SIMON, C. W. Arrangement of instruments, the distance between instruments and the position of instrument pointers as determinants of performance in an eye-hand coordination task. *USAF WADC tech. Rep.*, 1952 (Feb.), No. 5832.
- GUILFORD, J. P. *Fundamental statistics in psychology and education*. New York: McGraw-Hill, 1956.
- JOHNSGARD, K. W. Check-reading as a function of pointer symmetry and uniform alignment. *J. appl. Psychol.*, 1953, 37, 407-411.
- MORLEY, N. J., & SUFFIELD, N. C. The effect of pointer position on the speed and accuracy of check-reading groups of dials. *Naval Motion Stud. Unit Report*, Great Britain, 1951 (Oct.), No. 46.
- MURRELL, K. F. H. The use and arrangement of dials. *Instrum. Pract.*, 1952, 6, 520-526.
- WARRICK, M. J., & GREYER, W. F. The effect of pointer alignment on check-reading of engine instrument panels. *USAF AMC Memo. Rep.*, 1948, No. MCREXD-694-17.
- WHITE, W. J. Effect of dial diameter on ocular movements and speed and accuracy of check-reading groups of simulated engine instruments. *USAF AMC tech. Rep.*, 1949, No. 5826.
- WHITE, W. J. Effect of pointer design and pointer alignment on the speed and accuracy of reading groups of simulated engine instruments. *USAF AML tech. Rep.*, 1951, No. 6014.
- WHITE, W. J., WARRICK, M. J., & GREYER, W. F. Instrument Reading III: Check-reading of instrument groups. *J. appl. Psychol.*, 1953, 37, 302-307.

(Received May 27, 1963)

COMBINING CHECK-READING ACCURACY AND QUANTITATIVE INFORMATION IN A SPACE-SAVING DISPLAY

SIDNEY G. DASHEVSKY¹

United States Army Human Engineering Laboratories, Aberdeen Proving Ground, Maryland

An experiment was conducted to investigate the possibilities that a check-reading display could be designed to yield quantitative information, and that a more compact format could be used, saving space while preserving information content. The commonly used 4×4 display matrix and 0.50 sec. exposure time were used. Application of Gestalt principles proved advantageous. In an earlier study the principle of continuity was found highly efficient for qualitative readout. In the present case, the principle of similarity of form was found to allow quantitative readout with little or no loss of check-reading efficiencies, which ranged between 94% and 98% correct detection of errors. Compressing the display by use of semicircular, rather than circular, dials improved performance even beyond its earlier, almost perfect, level.

When numbers of dials are functionally grouped and must be intermittently but quickly monitored, we have a task generally referred to as check reading. Numerous prior studies have demonstrated that arranging dials in a matrix so that their pointers form a pattern in normal operation significantly aids detection of deviation. A bibliography of check-reading research is contained in Dashevsky's (1964) paper. All studies to date have treated check reading as a "go-no-go" situation in which displays are designed solely for the presentation of qualitative information. Other similarities mark the course of research in this area: exposure times of 0.50 or 0.75 second have been most common (there being agreement that this closely approximates the period for which a pilot may safely direct his attention from the windshield); dials have been arranged in a 4×4 matrix and have simulated meters having full circular faces.

The present study is addressed to the problems of: (a) whether dials can be made to yield quantitative information while permitting at least 90% correct detection of errors in the check-reading situation, and (b) whether a shorter scale, and consequently smaller display can be used in check reading. The first question is raised by the fact that meters are not normally designed nor intended for the presentation of solely qualitative data. Rather, they operate through a

range. Where qualitative information is all that is desired, task demands are better served by use of warning lamps or annunciators. This may account for engineers often treating the results of studies on dial check reading as being interesting, but unrelated to reality. The second point was suggested by the observation that in the usual situation employing circular dials for check reading, fully 180 degrees of the dial face is incapable of yielding information.

In a prior study (Dashevsky, 1964; Dashevsky & Glucksberg, 1963) Gestalt theory provided a productive approach to the perceptual problems involved in display organization. The authors took advantage of the continuity principle to organize the display into a tighter configuration (errors showing up as breaks in an otherwise continuous display). With this change it proved possible to raise accuracy in a check-reading situation from 85% to 98%.

In seeking a cognate principle by which displays could be structured to provide both check-reading and quantitative information, the principle of similarity of form seemed appropriate. In this case, errors would appear as dissimilar figures against a homogeneous ground. Since meters almost always operate through a normal range, with the dial face being divided approximately into thirds (below normal, normal, and above normal), it seemed feasible that coloring the normal range a medium green would provide lowered

¹ Now at the University of Rochester.

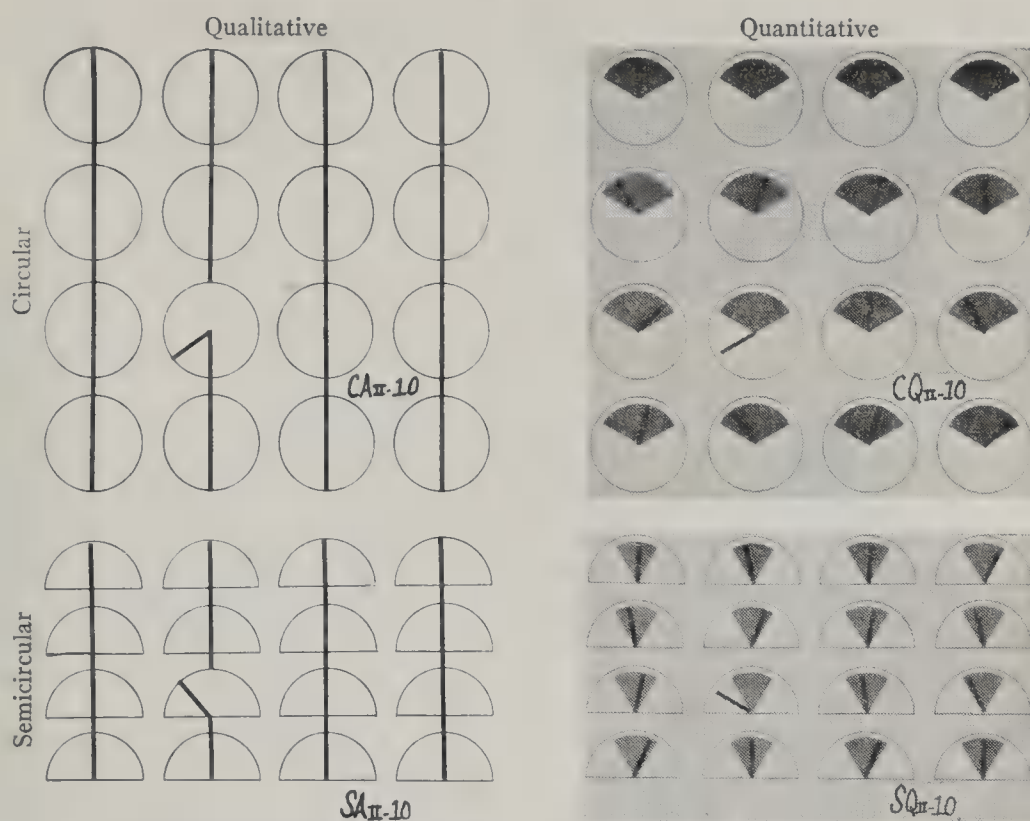


FIG. 1. The aligned (solely qualitative) display with one designed for quantitative readout, and full circular and semicircular dials.

contrast with pointers within that range, causing deviant pointers to stand out sharply against a white background. The green arcs would then form ground against which deviations would assume a figural aspect.

The object of reducing spatial requirements of displays was approached by using dials of identical radius having semicircular rather than circular faces.

The experimental design, then, is a factorial one comparing: (a) the aligned (solely qualitative) display with one designed for quantitative readout, and (b) full circular and semicircular dials (see Figure 1).

METHOD

Apparatus

Displays were simulated by 35 millimeter slides projected by a Tower Model Number 9880, automatic slide projector, projecting a 39-inch square image. Subjects (Ss) sat 7-10 feet from the screen and used response sheets on each of which 10 4×4 representations of displays were printed. Response forms were held in clipboards illuminated by shielded flashlight bulbs. Exposure was controlled by an Ilex Number 4 shutter fitted over the projector lens.

Subjects

Forty officer, enlisted and civilian personnel of the Human Engineering Laboratories were used. All were pretested on a task similar to the experimental one and assigned to one of four groups on the basis of pretest score. Due to marked skewness of score distributions, it was not possible to match Ss, so assignment was made solely with the object of equating mean pretest error scores of the four experimental groups. *T* tests between groups' mean ages and educational levels were not significant. It should be noted that since the *t* test for uncorrelated means had to be used (Ss not having been paired), the possibility exists that the *t* test for groups having correlated means would reveal one or more significant differences between mean age or educational levels. Inspection of the data, however, led the author to believe that such difference would have no systematic effect upon the perceptual task involved.

Procedure

At least 2 days after pretest, Ss were assigned to the four groups. Each group was shown 10 (unscored) training slides and two test sets of 10 slides each. All slides simulated 4×4 displays and were exposed for 0.5 second. Serial presentation of deviations, location in the display of deviant dials and direction of deviation were randomly selected and were identical for the four experimental series.

Deviations were of constant magnitude. Color slides (35 millimeter) were prepared by photographing acetate overlays placed over simulated display panels. All slides were photographed, developed, and mounted together to control background density. Errors were scored if Ss failed to detect an error or reported a nonexistent one (1 error) or ascribed a deviation to the wrong dial after successful detection ($\frac{1}{2}$ error). Erasures were counted as errors or half errors when they appeared. No more than one dial was deviant on any slide.

RESULTS AND DISCUSSION

The circular-aligned (qualitative) display was chosen as a standard because it had yielded the most efficient error detection in the prior study. Accuracy scores for this display were virtually identical in the earlier and present studies. Percentage correct scores are presented in Table 1. It will be seen that use of the quantitative display results in a slight loss of efficiency when circular dials are used and considerably less with semicircular dials. It is also apparent that use of semicircular dial faces noticeably improved accuracy.

The 2 × 2 factorial design was chosen to permit analysis of variance. However, the virtual nonexistence of errors (and so, of variance) effectively precluded use of this statistic. Nor could nonparametric statistics be validly applied since they rely upon one of three assumptions: independence of sampling, symmetrical distributions within cells, or matching of pairs. This study violates all three: group means (but *not* individual members) had been matched, all cells showed truncation, and pairs had not been matched.

A test for significance of differences was,

TABLE 1
PERCENTAGE CORRECT RESPONSES

	Qualitative	Quantitative
Circular	97.75	94.25
Semicircular	99.75	97.75

TABLE 2
NUMBER OF PERFECT SCORES

	Qualitative	Quantitative
Circular	5	3
Semicircular	9	6

Note.—N per cell = 10.

however, felt to be desirable and with the above factors in mind, a chi-square test was performed. This seemed the best choice, for while the violation of the other two assumptions was indisputable, the lack of independence of sampling was by no means conclusive. This was lent credence by the observation that Ss' pretest and experimental scores appeared to be only moderately correlated. The chi-square test, using Yates' correction for continuity, was applied to the frequencies of perfect scores in each cell (see Table 2) and chi-square was found to be 10.99 ($p < .001$).

We may conclude that the qualitative display was significantly superior to the quantitative, and the semicircular to the circular, but efficiencies are of an order that the least efficient display will yield highly satisfactory performance should its use be indicated.

Since it was not possible to perform an exact test for interaction effects, further research seems indicated. This could be accomplished by replicating the study using exposure times of 0.5 or even 0.1 second. The number of errors would presumably increase and be normally distributed, allowing an analysis of variance to be performed.

REFERENCES

DASHEVSKY, S. G. Check-reading accuracy as a function of pointer alignment, patterning, and viewing angle. *J. appl. Psychol.*, 1964, 48, 344-347.
DASHEVSKY, S. G., & GLUCKSBERG, S. A method for increasing efficiency of dial check-reading. *USA Ordn. Hum. Engng. Lab. tech. Memo.*, 1963, No. 6-63.

(Received September 30, 1963)

VALIDATION OF A MULTIPLE-ASSESSMENT PROCEDURE FOR MANAGERIAL PERSONNEL

PAUL A. ALBRECHT

Claremont Men's College and

Edward Glaser & Associates

EDWARD M. GLASER

Edward Glaser & Associates

AND JOHN MARKS

Mental Health Research Institute, Ft. Steilacoom, Washington

A multiple-assessment procedure—personal history form, intensive interview, 2 objective intellectual aptitude tests, a sentence-completion test, and a human relations problems test—was used to predict the performance of 31 industrial managers all having a similar job assignment. Predictions were made on the basis of a global, nonactuarial analysis of these objective and subjective data. 4 sets of criterion judgments were obtained on 4 variables—3 different sets of rankings and 1 set of ratings. A multitrait-multimethod matrix was used in the analysis of the intercorrelations. 9 of the 12 validity coefficients involving ranking-type criteria were statistically significant. Of the 4 coefficients involving rating-type criteria, none was significant.

Psychologically based managerial assessment procedures, although frequently highly prized and rather widely used, still lack adequate validation. In addition, the methodology itself seems currently to be the subject of a healthy, if somewhat agonizing, reappraisal. Instead of having proceeded logically from development to application, the managerial assessment field (by way of borrowing of techniques) can almost be said to have proceeded from application to development. The current and slowly-growing emphasis on validation, of which this study is a part, is a mature, if time-inverted trend.

One aspect of the methodological reappraisal involves a lively and at times acrimonious dispute over methods of securing and analyzing predictive data which has flared on two major fronts. The first front finds the psychometric and personnel selection professionals defending themselves against an attack from journalists on selection testing in general and personality testing in particular. Gross's recent book (1962) is an example.

The second front, somewhat paradoxically, finds the psychometric group (actuarial branch) on the offensive against the clinical prediction group. Meehl's (1954) able opposition of the actuarial and clinical approaches, rather to the disparagement of the

latter, has served to highlight, perhaps also to crystallize, this issue. The elaborate Kelly-Fiske (1951) study comparing the efficacy of various approaches to the prediction of the performance of clinical psychologists has done little to allay concern in the clinical prediction camp. Recently, however, a counterattack has developed on this front exemplified, for example, by Holt's (1958) charge, supported by some evidence, that limitations in clinical prediction studies cited by Meehl were more an indictment of clinicians' research interests and skill than a real test of the clinical method.

While the controversy continues, much managerial and executive selection activity is nevertheless actually going on and in fact has been in process since World War II. A surprising number of firms, as well as individuals, have used a variegated procedure (lengthy interview, personal history form, a few objective tests, and a few of the less costly projective tests) which for want of a better term will be here referred to as multiple assessment. While the practitioners are typically clinical in their orientation, the increased use of personal history and psychometric data shifts the emphasis somewhat toward the objectifying of inference even though the final decision is based on clinical rather than actuarial procedures.

What is the current evidence on the validity of such procedures? In regard to the individual techniques which make up the battery of procedures, reviews of the literature on the interview such as by Wagner (1949), for example, are quite critical of its typical application. Others have shown some improvement in interview validity when more standardized and specific objectives and procedures are involved (McMurry, 1947; Yonge, 1956). Psychometric approaches, at least to this point, seem to tail off in their validity as they approach the supervisory and management selection area (Ghiselli & Barthol, 1953).

Direct tests of the multiple-assessment procedure are few, but some cautiously supportive evidence has been emerging. A series of studies coming from the Personnel Research Services group at Western Reserve are directly relevant. Hilton and his co-authors (Hilton, Bolin, Parker, Taylor, & Walker, 1955) showed some promising results in a preliminary study comparing psychologists' ratings based on assessment reports with mailed criterion ratings from the companies in which these managers were employed. In a later series of studies which presented evidence on multiple-assessment validity as well as on the validity of the several components, some validity was again reported for the total multiple-assessment procedure. Correlations ranged from $-.05$ to $.46$ (J. T. Campbell, Otis, Liske, & Prien, 1962).

Turning to the individual-component procedures in the above study, the interview showed slight but not strong validity (Prien, 1962). The projective instruments showed questionable results (Hogue, Otis, & Prien, 1962). Objective psychometric devices did not show strong relationships to the criterion either (J. T. Campbell et al., 1962). But predictor ratings based on the psychometric data showed the best of the component validities (Huse, 1962).

Some further positive evidence on a multiple-assessment procedure is provided by Trankel's study (1959) of a slightly different type of selection, that of airline pilots. Results showed that a multiple-assessment approach resulted in higher validity coefficients

than standardized tests alone. In a 2-year follow-up study of assessment procedures first carried out for the selection of first-line supervisors, Handyside and Duncan (1954) reported on a study conducted in England in which rather surprising validities are shown.

This brief survey shows that the evidence, while far from conclusive, does suggest some possible validity for executive selection of a multiple- or a semiclassical-assessment procedure. Presumably one can assume that, as usual, studies indicating negative evidence are less apt to be found in print. The following study, with improvement on the criterion side over typical studies, is offered as additional evidence on multiple assessment.

EXPERIMENTAL DESIGN

The subjects (Ss) were managers of a national corporation, each occupying the same recently created middle-management job, that of district marketing manager. Each was located in a separate geographical district. These jobs probably received somewhat different phrasing in the several districts, but the official job duties and responsibilities were identical. The relative equality of jobs made possible certain rather unique criterion development. The managers were all promoted from within the company. They ranged in age from 29 to 49. The median age was 36. The median length of service with the company was 11 years, 8 months.

The assessment procedures were not used in the selection process in any way and copies of the reports were not given to the company during the evaluation period. However, each *S* was shown his own report and given opportunity to discuss it briefly in the interest of the humane reduction of tension and because the corporation wished to gain this individual development advantage from its extensive support of the project.

The firm had not used this assessment procedure previously and top management agreed to gather the criterion information and support the project in order to try out its validity in this setting. The practical exigencies of arranging the study resulted in the assessments being done soon after, rather than prior to, the appointment of the incumbents. Almost immediately after being selected, the managers filled out the personal history forms and the tests, with the exception of the timed Problems test. Nearly all the interviews themselves were conducted within the first several months. Some, including two Ss who were replacements for managers who were promoted, were done during subsequent months. The delay in finishing some of the assessment interviews had no apparent contamination effect, as will be seen later.

Two staff members of the firm of consultants did all but two of the interviews, a third staff member being responsible for the remaining two. One of these two principal staff members had some prior contact with the firm; the other had none, except during the assessment procedure itself. The contact of the first staff member was very minimal during the year evaluation period, being almost entirely confined to the administration of the validation program itself.

After the data on each man has been collected, an assessment report descriptive of each manager was written. The actual ranking and rating predictions were made near the end-of-the-year evaluation period but prior to the collection of the criterion data. Since the jobs were completely new to the company, it was felt that more experience with the position would be needed before meaningful criterion variables could be selected. Hence, they were not selected in the beginning.

The assessment battery itself included a rather extensive interview, involving in part an elaboration of data appearing on a personal history form which *S* had filled out prior to coming to the interview. These interviews typically took about 2 hours. In addition, *Ss* took a brief intellectual aptitude test, the Problems test, during the assessment interview. In advance of the interview situation, *S* had written responses to a sentence-completion test, a Human Relations Problems Test, and had taken the Watson-Glaser Critical Thinking Appraisal.

Prior to the final development of the criteria, it had been hoped that some objective indicators of performance could be procured—in addition to the more usual global evaluations. None was judged to be adequate. Objective indicators considered, such as district sales or cost performance, were found to be subject to too many other factors in addition to the manager's performance. His responsibility in some of these areas was more of a staff than a direct-line nature.

In view of the many persistent difficulties with supervisory ratings, it was decided to emphasize rankings made by regional superiors and by peers. This was possible since the 31 districts were organized into 3 regions. Within each region, the regional general manager (2 levels removed), the regional marketing manager (a staff man having functional responsibility for performance), and the marketing manager peers were asked to rank the managers in the region from highest to lowest on 4 variables. A representative of the headquarters personnel office visited each region, explaining the ranking method and supervising the procedural aspect of making the rankings.

In addition, the district manager (the immediate supervisor) was asked to rate his man on the same four variables, except that in his case an absolute rating scale, rather than a ranking, was used since he knew only his immediate subordinate's performance. The peers, although working in separate districts, had spent considerable time in mutual

discussion and planning with their counterparts in the region in regard to this new position. Hence they had considerable knowledge of certain aspects of each other's performance. These several sets of judgments, in addition to providing a wide range of criterion information, also made possible comparison among several methods as well as several loci of managerial performance evaluation.

Three areas of performance were finally selected for evaluation: (a) forecasting and budgeting effectiveness, (b) sales performance, and (c) effectiveness in interpersonal relationships. The fourth variable was overall performance.

The consultants similarly made predictor rankings on these four variables for each of the three regions. One of the staff members did all the assessments in Region A; a second staff member did those in Region B. Interviews in Region C were of practical necessity divided among all three staff members. In this case, each estimated rankings for his *Ss*, and a combined ranking was then developed by one of the staff members from these provisional rankings and from the assessment reports. This proved to be a difficult task. A subsequent check of the data showed the value of a consistent and unitary frame of reference for prediction such as was present in Regions A and B.

In order to gain the statistical advantage of larger sample size, the assumption was made that the three regional groups were equivalent in the caliber of their marketing managers, and the rankings were converted to a common numerical base and recombined into common rankings for the entire group.

RESULTS

The major results of the data analysis appear in Table 1. The data gathered in this study can be regarded as a set of five method variables and four trait variables (D. T. Campbell & Fiske, 1959). The five method variables include the four sets of criterion judgments and the set of predictor judgments. The four trait variables are the performance dimensions. Among the four groups making the criterion judgments there was undoubtedly some communication, presumably more among peers and among superiors than between these two subgroups. The predictor rankings were made prior to the criterion rankings and, of course, independently of them.

In Table 1, the italicized values represent the validity coefficients. They are the same traits measured by different methods. The triangular blocks of correlations bordering the diagonal represent the heterotrait-monomethod correlations.

TABLE 1
PRODUCT-MOMENT INTERCORRELATIONS OF PERFORMANCE RANKINGS AND PREDICTIONS^a

	District managers				Regional general managers				Regional marketing managers				Peers				Consultants' predictions			
	F	S	I	O	F	S	I	O	F	S	I	O	F	S	I	O	F	S	I	O
District managers																				
Forecasting	35	49	71		33	11	35	31	29	-04	29	20	50	-09	26	30	-04	-26	-02	-09
Sales		17	55		27	50	30	30	-07	32	21	24	18	36	09	19	00	08	-04	04
Interpersonal			71		36	30	38	31	27	00	34	18	33	04	36	34	13	-02	12	-01
Overall					37	24	36	36	20	09	35	23	51	07	24	37	11	-08	-01	07
Regional general managers																				
Forecasting					54	51	90		66	34	44	54	37	24	30	42	35	11	02	26
Sales						32	63		14	57	31	39	-08	40	-02	08	10	35	14	14
Interpersonal							68		23	38	62	50	42	21	42	42	17	30	28	24
Overall									45	52	58	67	37	36	36	48	28	28	19	34
Regional marketing managers																				
Forecasting									04	33	42		47	-07	11	26	41	-07	-17	21
Sales										67	79		-08	74	19	36	13	52	16	30
Interpersonal											85		34	54	50	66	37	40	25	37
Overall													22	60	32	54	18	29	01	27
Peers																				
Forecasting													-04		53	60	44	01	22	39
Sales															41	62	14	61	29	40
Interpersonal																87	29	27	54	38
Overall																	42	42	47	56
Consultants' predictions																				
Forecasting																				
Sales																				
Interpersonal																				
Overall																				

Note.—N = 31. Forecasting = F, Sales = S, Interpersonal = I, Overall = O; r = .30, p < .05 (one-tailed), r = .42, p < .01 (one-tailed).
a Decimal points omitted.

TABLE 2
MEDIAN TRAIT INTERCORRELATIONS
FOR EACH SET OF JUDGES

Self-trait intercorrelation	Number <i>r</i> 's	Median <i>r</i> 's ^a
District manager	3	35
Regional general manager	3	51
Regional marketing manager	3	33
Peer	3	41
Consultant	3	45
Combined	15	44

■ Decimal points omitted.

Arranging the data in this form permits the answering of some basic questions. Before considering these questions, it might be noted further that some median correlations have been extracted from Table 1 and are presented in Tables 2 and 4. Because the ratings of overall performance were intended as a kind of combination of the other three ratings, correlations involving overall ratings were not included in computing these medians with the exception of the one case representing the median validity coefficient of the overall ratings themselves.

First, one may ask, are the trait ratings independent? The median self-intercorrelations on the traits presented in Table 2 show that this is not the case. Persons rated high on one variable tend to be rated high on another.

TABLE 3
RANK-ORDER INTERRELATIONSHIP BETWEEN
CRITERION DIMENSIONS

Criterion dimension	Regional general manager	Regional marketing manager	Peer	Con- sultant
Forecasting-budgeting versus sales	55	07	-04	46
Forecasting-budgeting versus interpersonal relationships	50	41	52	47
Forecasting-budgeting versus overall	90	44	59	84
Sales versus inter- personal relationships	32	67	41	68
Sales versus overall	64	79	61	73
Interpersonal relation- ships versus overall	68	85	87	67

Note.—Decimals omitted.

This can be thought of as halo effect or as an expression of a fact of life that good traits tend to come together in the same package. Operationally in this study both explanations amount to the same thing.

Of the raters, the regional general managers show the least independence of ratings, the regional marketing managers the most. Yet these self-correlations are not high enough

TABLE 4

VALIDITY AND HETEROTRAIT-HETEROMETHOD
CORRELATIONS

	Number <i>r</i> 's	Median <i>r</i> 's ^a
Validity correlations (same traits, different methods)		
Forecasting—budgeting	10	39
Sales	10	45
Interpersonal relationships	10	37
Overall effectiveness	10	37
Combined	40	38
Heterotrait-heteromethod correlations (different traits, different methods)		
District Manager × Regional General Manager	6	30
District Manager × Regional Marketing Manager	6	11
District Manager × Peers	6	14
District Manager × Consultants	6	-02
Regional General Manager × Re- gional Marketing Manager	6	32
Regional General Manager × Peers	6	22
Regional General Manager × Consultants	6	13
Regional Marketing Manager × Peers	6	15
Regional Marketing Manager × Consultants	6	15
Peers × Consultants	6	25
Combined	60	17

■ Decimals omitted.

to justify the conclusions that the three traits are identical. For all of these median self-correlations, 70% or more of the variance is independent.

A somewhat different method of analysis than these median self-intercorrelations appears in Table 3. Here rank-order correlations, necessarily excluding the district manager criterion judgments not given in rank

form but now including the overall dimensions, have some interesting applications.

The forecasting and budgeting versus sales relationship is the lowest. This partially substantiates what had been an a priori characterization of the forecasting and budgeting function as involving a more individualized, analytic, and intellectual function as compared to the personal contact and supervisory components of the sales function. It is interesting to note that this separation is more clear to those closest to the day-to-day activities (peers and regional marketing managers) than to those more removed (regional general managers and consultants). There may, of course, be other reasons for this apparent difference than distance from job functioning.

It is also evident that regional general managers and consultants show a closer relationship between forecasting and budgeting and overall evaluations than do the peers and regional marketing managers. Some differences in the assignment of priorities in the job definitions are obvious here. In general, the dimension of sales performance shows the most independence from the others.

A second question, in addition to the previous one having to do with independence of judgments, may now be asked. Are the ratings valid? A partial answer to this question is seen in the median intercorrelations given in Table 4, which were extracted from the heterotrait-heteromethod matrix in Table 1. The question being asked in this method of analysis is whether measures of the same trait by different methods are more closely

related than measures of different traits by different methods. This obvious requirement of a valid set of criterion judgments is frequently not met.

In general, it will be seen that this validity condition is met. The median combined validity correlation is .38, while the median combined heterocorrelation is .17. Again, the sales variable seems to be the one which shows the highest intermethod correlation. The two sizable heteromedians involve the regional general manager judgments which were, as noted before, the ones which showed the least independence among the various traits. This lack of independence of the trait ratings would tend to inflate the heterotrait cells.

Third, it can be asked, and this was the main point of the study, did the multiple-assessment judgments of the consultants predict how well these marketing managers later would do their jobs? One answer to this can be seen in the four submatrices in the last column of Table 1. Of 16 possible predictive validity coefficients, 9 are significant at the .05 level or better. When the validity coefficients between predictor and district manager judgments which are based on ratings are removed, 9 of the 12 validity coefficients with the ranking criterion method are significant.

The consultants seem best able to predict performance in sales and forecasting-budgeting areas and are somewhat less accurate in predicting interpersonal relationships. In each of the four categories, their greatest agreement was with the peer ratings, which may have been more reliable since they represented the means of the ranking of a number of judges, rather than individual ranks as were the regional general manager and regional marketing manager measures.

One way of making the criterion more reliable is to combine all the rankings, regional general managers, regional marketing managers, and peers into one combined criterion. To this end, the rankings from these 3 methods were summed. The correlation of this combined criterion with the consultants' predictor rankings is shown in Table 5. Here the results are more consistent and impressive. All of the diagonal correlations are

TABLE 5

INTERCORRELATIONS OF CONSULTANTS' PREDICTION
RANKINGS WITH THE COMPOSITE CRITERION

Consultants' rankings	Composite criterion			
	Fore- casting	Sales	Inter- personal	Overall
Forecasting	49	15	33	34
Sales	02	58	39	39
Interpersonal	03	24	43	27
Overall	35	33	40	46

Note.—Decimals omitted.

significant beyond the .01 level, and none of the heterotrait correlations are as large as the monotrait ones. Apparently the multiple-assessment predictions were able to distinguish between the traits successfully in their predictions of performance.

The results given in the last column of Table 1 show that the consultants' rankings were not significantly related to the district-manager ratings. It will be recalled that each district manager rated his own marketing manager on an absolute scale rather than making relative rankings over a number of men. Additional analysis corroborates this finding. At the time the consultants made their predictor rankings, they also made additional ratings for each man upon the absolute scale. In Table 6 are shown the intercorrelations. The consultants failed to predict the absolute ratings.

A general question in regard to these predictions may be raised on the basis of the validation-study design. It may be asked whether the consultants, since they saw some of the marketing managers after they had started their jobs, were not able in the interview to pick some clues of how they were doing. In such a case the consultants' judgments might contain elements of concurrent rather than predictive validity.

A partial test of this was made. The overall performance rankings for each region from the combined criterion were compared with the consultants' rankings for that region. The deviation of the two rankings is an index of how far the consultants' rankings went astray

TABLE 6

INTERCORRELATIONS OF CONSULTANTS' PREDICTION RATINGS WITH DISTRICT MANAGER RATINGS

Consultants' ratings	District manager ratings			
	Forecasting	Sales	Interpersonal	Overall
Forecasting	.07	.01	.21	.17
Sales	-.23	.23	-.02	.00
Interpersonal	-.03	.00	.19	.12
Overall	-.06	.16	.07	.09

Note.—Decimals omitted.

TABLE 7

INTERCORRELATIONS OF TESTS WITH CRITERION AND PREDICTOR RANKINGS

Performance dimensions	Tests	
	Problems test	Critical thinking test
Forecasting-budgeting		
Combined criterion	.41	.30
Predictor's ranking	.53	.48
Sales		
Combined criterion	-.07	.12
Predictor's ranking	-.16	.29
Interpersonal relationships		
Combined criterion	.18	.18
Predictor's ranking	-.15	.06
Overall effectiveness		
Combined criterion	.23	.24
Predictor's ranking	.25	.47

Note.—Decimals omitted.

of the mark. If there was tendency for the consultant to pick up clues in the interview as to how the manager was doing, the later interviews, occurring several months after the first ones, should show the smallest deviation. A correlation was run between the deviations and the months from the beginning at which the particular interview was made. The resulting correlation was an insignificant .03.

In view of the continuing attention to the comparative validities of psychometric versus multiple-assessment or clinical approaches, the psychometric data are of interest. The two objective psychometric instruments used were both measures of intellectual functions: the Problems test (brief general mental ability test) and the Watson-Glaser Critical Thinking Appraisal (test of functional effectiveness in applied reasoning). There are no data in this particular study on objective personality tests. The sentence-completion test and the Human Relations Problems Test in this battery were not analyzed independently or quantitatively. They were treated qualitatively in the same fashion as were the interview and biographical data.

The 2 objective tests correlated .54 with each other. Of 32 possible criterion correlations for the 2 tests (each correlated with the 4 traits rated or ranked by the 4 judge groups) only 4 are significant. Three of these 4 occur in relation to predictor and criterion judgments on the forecasting and budgeting performance dimension. The fourth significant correlation is between the Critical Thinking Appraisal and the predictors' ranking of overall effectiveness. It will be remembered, by comparison, that the multiple-assessment rankings showed 9 significant out of 16 possible criterion correlations. The correlations between the objective tests and criteria are given in Table 7 which follows.

DISCUSSION

Several important and interesting issues are raised by the foregoing results. Of most basic import is the finding that the use of multiple-assessment procedures did result in generally valid predictions. Perhaps the currently fashionable attacks on multiple-assessment and/or clinical procedures are premature and overgeneralized. Holt's previously mentioned statement (1958), that much of the negative evidence on clinical methods of prediction was really gleaned from poorly designed and executed studies and so was at least partially misleading, is born out by recent studies, including this one.

The generally valid predictions made in this study, it should be noted, were made when the sample was restricted to men who had already been selected by their company for the job. None of the candidates passed over in the selection process appeared in this population.

It is, of course, generally expected that elaborate and expensive assessments of this sort will be used with relatively restricted populations of candidates, perhaps including only those who have survived preliminary screening. But limiting the sample to those already selected presents a more severe than usual test. Also it should be noted that the predictions were of a rather rigorous and differentiated nature in that they involved ranking Ss rather than merely establishing a pass-fail criterion.

Of further interest, in view of frequent current discussions, is the finding that the multiple-assessment procedure resulted in more valid predictions than did the psychometric tests. To be sure, the tests in this battery were limited to those of intellectual aptitude and performance. Furthermore, the tests did show some significant correlation with the performance dimension of forecasting and budgeting, which on an a priori basis would seem to be most heavily weighted with intellectual-performance factors. But the tests did not provide significant correlations with other criterion variables, including that of overall performance.

Looked at in one way, it might be said that what is shown here is that an appropriate test should be selected or devised for a given dimension. Additional types of tests, e.g., objective personality tests, would perhaps have predicted other dimensions. When there are many similar positions in a given job category or when frequently recurring dimensions can be found in even apparently dissimilar jobs, such psychometric approaches seem worth pursuing. Enthusiasm for this approach at this time, however, should be tempered by awareness of the technical difficulties with objective personality tests, the other major type of psychometric instrument. Currently available objective personality tests have shown many limitations in the prediction of managerial success—a fact which led to the decision not to include them in this multiple-assessment battery.

However, looked at in another way, the findings on the comparative validity of multiple-assessment and psychometric predictions tend to demonstrate that global or clinical procedures can "tailor-make" a prediction to a job-performance class which is relatively unique. This, of course, is one of the advantages nearly always claimed for such procedures. Many managerial jobs represent a rather unique combination of performance dimensions. This is particularly true when one takes into account, in addition to literal job duties, the organizational and interpersonal climate in which a manager is asked to function.

In regard to another aspect of the results, the inability of the consultants to predict the district managers' performance ratings is of considerable interest. The fact that there were three different sets of rankings, in addition to these district manager ratings, and that success was shown in the prediction of the sets of rankings, lends considerable credence to the conclusion that the difficulty in the district manager case was due primarily to the rating procedure by means of which these particular judgments were made. Mere multiplication of judges does not explain the results since there were also 31 judges responsible for the peer rankings, which were predicted with the highest accuracy of any set of criteria.

Several possible and more or less common explanations can be made. In the case of the district managers, each *S* was judged by only one person, leading to less stability in the criterion as compared with the multijudge peer-ranking procedure. The rating-form procedure, in contrast to ranking, adds the uncertainty of interpretation of a scale with the possibility for central tendency or leniency biases to enter the picture. Since none of these district managers was known to the predictors, the tailor-making possibility of the multiple-assessment procedure was not brought to bear in estimating the potentialities of *S* in the light of knowledge of his superior's expectations.

One other fortuitous factor in this particular case resulted from the fact that the job category was new, and therefore common or even stereotyped expectancies on the part of the district managers presumably had not yet had time to form.

Nevertheless, the frequency with which this multijudge rating procedure is used in validation studies makes the special difficulty encountered with it in this study of interest. The finding relates to recent comments in the literature to the effect that the manner in which the criterion is derived and analyzed may be a very significant factor in the results of validation studies of predictor devices (Bass, 1962; Ghiselli & Haire, 1960).

CONCLUSIONS

1. Predictions based on a multiple-assessment procedure were significantly related to criterion rankings of managerial performance made by two sets of superiors and by a set of peers. Nine of the 12 validity coefficients were statistically significant at the .05 level or better.

2. A fourth set of judges used an absolute rating scale instead of the ranking method. None of the four validity coefficients involving multiple-assessment predictions of these rating criteria was significant.

3. The two objective tests of intellectual functions showed some statistically significant relations to the forecasting and budgeting criterion variable and to the predictor judgments on the same variable. But they did not in general predict performance as well as the multiple-assessment procedure.

REFERENCES

- BASS, B. H. Further evidence on the dynamic character of criteria. *Personnel Psychol.*, 1962, 13, 93-97.
- CAMPBELL, D. T., & FISKE, D. W. Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychol. Bull.*, 1959, 56, 81-105.
- CAMPBELL, J. T., OTIS, J. L., LISKE, R. E., & PRIEN, E. P. Assessments of higher-level personnel: II. Validity of the overall assessment process. *Personnel Psychol.*, 1962, 15, 63-74.
- GHISELLI, E. E., & BARTHOL, R. P. The validity of personality inventories in the selection of employees. *J. appl. Psychol.*, 1953, 37, 18-20.
- GHISELLI, E. E., & HAIRE, M. The validation of selection tests in the light of the dynamic character of criterion. *Personnel Psychol.*, 1960, 13, 225-231.
- GROSS, M. L. *The brain watchers*. New York: Random House, 1962.
- HANDYSIDE, J. D., & DUNCAN, D. C. Four years—a follow-up of an experiment in selecting supervisors. *Occup. Psychol.*, 1954, 28, 9-23.
- HILTON, A. C., BOLIN, S. F., PARKER, J. W., TAYLOR, E. K., & WALKER, W. B. The validity of personnel assessment by professional psychologists. *J. appl. Psychol.*, 1955, 39, 287-293.
- HOGUE, J. P., OTIS, J. L., & PRIEN, E. P. Assessments of higher-level personnel: VI. Validity of predictions based on projective techniques. *Personnel Psychol.*, 1962, 15, 335-344.
- HOLT, R. R. Clinical and statistical prediction: A reformulation and some new data. *J. abnorm. soc. Psychol.*, 1958, 56, 1-12.
- HUSE, E. R. Assessments of higher-level personnel: IV. The validity of assessment techniques based

- on systematically varied information. *Personnel Psychol.*, 1962, **15**, 195-205.
- KELLY, E. L., & FISKE, D. *The prediction of performance in clinical psychology*. Ann Arbor: Univer. Michigan Press, 1951.
- McMURRY, R. N. Validating the patterned interview. *Personnel*, 1947, **23**, 2-11.
- MEEHL, P. E. *Clinical versus statistical prediction*. Minneapolis: Univer. Minnesota Press, 1954.
- PRIEN, E. P. Assessments of higher-level personnel: V. An analysis of interviewers' predictions of job performance. *Personnel Psychol.*, 1962, **15**, 319-334.
- TRANKEL, A. The psychologist as an instrument of prediction. *J. appl. Psychol.*, 1959, **43**, 170-175.
- WAGNER, R. The employment interview: A clinical summary. *Personnel Psychol.*, 1949, **2**, 17-46.
- YONGE, K. A. The value of the interview: An orientation and pilot study. *J. appl. Psychol.*, 1956, **40**, 25-31.
- (Early publication received January 14, 1964)

SENSORY-FEEDBACK ANALYSIS OF BEHAVIOR IN STEREOTELEVISED VISUAL FIELDS

KARL U. SMITH¹ AND JOHN D. GOULD²

University of Wisconsin

This research has devised special methods of 3-dimensional television in order to explore several main problems of visual-feedback control of behavior: (a) evaluate techniques of achieving remote mobile 3-dimensional vision; (b) analyze sensory-feedback control of behavior by means of 3-dimensional vision; (c) preliminarily evaluate the design problems of machines guided with remote stereotelevision and similar display channels. The research was divided into 2 main parts—an initial phase of equipment development and evaluation and a 2nd phase of controlled experiments. The initial experiments have been reported earlier. In the present sensory-feedback research, the results indicate that stereotelevised feedback has major visual limitations, which can be overcome in part by instrumental control and that different task patterns with 3-dimensional remote vision present specialized characteristics of sensory feedback, dynamic in nature and best assessed by direct visual-feedback research.

Research on stereotelevision has many scientific and technological implications. Basically, aside from its applied importance, its greatest significance in visual research is in providing flexible but precise experimental control of the binocular visual feedback of optically-controlled manual behavior. With such experimental control we can initiate the study of the real-time and real-space functions in three-dimensional vision and optically controlled motion (Smith, 1962, 1963; Smith & Smith, 1962).

The technological applications of stereotelevision include the optical design of remote control and manipulation systems for lethal and dangerous materials and for remotely guided anthropomorphous machines and space vehicles. Already, machines and devices of this nature employ substitute television eyes for

visual control and guidance, and in some cases they may be equipped with stereotelevision systems.

Three approaches to stereotelevision have been developed. Mengle (1958) has described a three-dimensional color separation system which we evaluated in a previous study (Gould & Smith, 1963). Mauro (1961) has reported an extensive study of optical accuracy of a three-dimensional color system also. In his system the images formed by two lens systems, separated by about 5 inches, are superimposed by means of a beam splitter so that both images fall on the aperture of a single image orthicon tube. This signal is passed through a filter drum containing Polaroid and color filters "to effect a color and stereo presentation for direct viewing with special Polaroid viewers." The physics of the system was described and tests of visual acuity reported with one subject (*S*).

In a series of six studies already reported (Gould & Smith, 1963) we have carried on developmental work in devising a binocular-separation stereotelevision system and in comparing visual efficiency in this system with that obtained in a color-separation technique. In the present research, a second series of systematic studies have been completed concerning three-dimensional sensory-feedback control of behavior. These studies demonstrate the possibility of using stereo-

¹ This experiment's funds came from the National Science Foundation for the study of sensory-feedback analysis of motion (Project NSF-7589) and from the National Institutes of Health for the investigation of geometric organization of human motion (Project 4469). The Graduate Research Committee, the University of Wisconsin, provided matching funds for the purchase of stereotelevision equipment and related motion analysis instrumentation. The Numerical Analysis Laboratory, the University of Wisconsin, aided in data processing. These studies were carried out in general cooperation with W. M. Smith, Dartmouth College.

² Public Health Fellow. At IBM Research Center, Yorktown Heights, New York.

television in sensory-feedback analysis of vision and also provide data about the organization of optic behavior with three-dimensional vision. The concepts related to the design of these specific studies will be given in connection with the description of the results of the experiments.

TECHNIQUE

A stereotelevision chain has been devised in which the monitor fields of the two cameras were completely separated and fused by means of prisms. In our original studies (Gould & Smith, 1963), this system was found to be free of some of the basic defects that have been found with color separation techniques. The apparatus used in the present work is shown in Figure 1. It consists of two RCA TK-201 cameras which are connected separately to two matched 8-inch Conrac Monitors (Model CN A8/C) mounted in a single rack adjacent to one another. A multiplex arrangement was devised so that the signal from each camera could be switched to either one of the two monitors. A stereoscopic prism hood was attached to the faces of the two monitors. The prisms in this hood could be adjusted for interpupillary width and for prism power independently in each eye in both the horizontal and vertical meridians.

EXPERIMENT I: OBJECT MANIPULATION WITH STEREOTELEVISION

This experiment was carried out to investigate the role of depth factors in the different component movements of object manipulation in assembly performance. The task situation illustrated in Figure 1 was employed. This consisted of a peg-board assembly task composed of a pin bin and assembly plate. The task of the *S* consisted of lifting a single pin (0.25 inch in diameter by 0.75 inch long) from a small bin and placing the pin in one of the holes in the assembly plate. The assembly plate contained eight rows of eight holes each.

Method

In order to measure the duration of the different component movements in the assembly performance, a four-channel electronic motion analyzer was attached to the assembly bin, the assembly plate, and *S*. This analyzer passes a current of about 100 microamperes through *S*. This subliminal current is used to activate relays and time clocks. When *S* touches the pins in the bin, the grasp relay of the motion analyzer closes and operates one of the clocks. As soon as the contact of the pin with the bin is broken by the action of *S* in lifting it, this grasp clock stops and the second or "loaded-travel" clock runs. This clock runs until the pin is brought into contact with the assembly plate. This contact automatically stops the loaded-transport clock and starts the

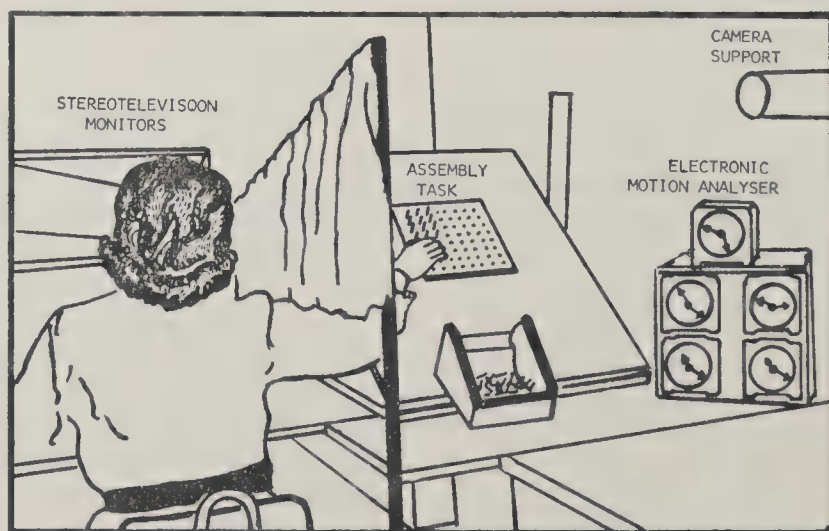


FIG. 1. The integration of stereotelevision and electronic motion-analysis instrumentation for study of object manipulation with three-dimensional television. (The end of the camera support is shown in the upper right of the diagram. The relays of the electronic motion analyzer are contained within the clock housing. The clocks used measure movement time in .01 second.)

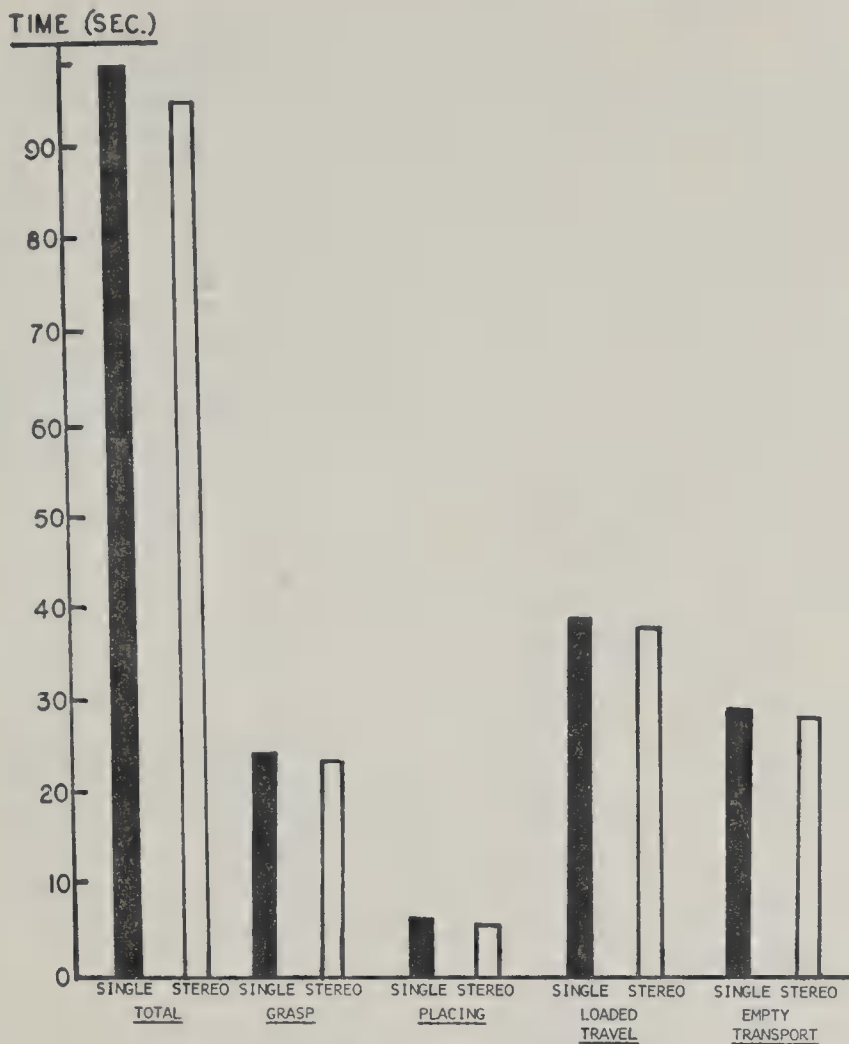


FIG. 2. The relative effect of single-camera and stereotelevision viewing upon different component movements in an assembly cycle.

"placing" clock. This third clock runs as long as S's hand maintains contact with the pin as it is inserted into the assembly plate. Once this contact is broken, the placing clock stops running and the fourth clock, the "empty-travel" clock, starts to run. This clock continues to run until contact is made again with the pins in the bin.

The assembly motion analyzer automatically summarizes the duration of the four separate component movements as the assembly cycle is repeated. When S places the last pin in the assembly plate, the pin in the last hole closes a "stop" relay that automatically stops the timing. The fifth clock shown on the motion analyzer gives the total time of all four component movements.

Twenty-four Ss served in this experiment. Each individual was given two trials under both the stereoscopic and the monocular viewing conditions. The order of conditions was counterbalanced. The time scores of the different component movements constituted the main data of the study. The measures for the second trial of each viewing condition were selected for data analysis. The Ss were instructed

to work as fast as possible. The two camera lenses were separated by a distance of 12.6 inches. The midline distance of the cameras to the workboard was 80 inches.

Results

The results of this study showed the mean total time for completion of the task for all Ss to be 97.93 seconds under the single camera condition and 94.90 seconds under the stereoscopic condition. In an initial analysis of the data, this difference was found not to be statistically significant. A more complete analysis of variance was therefore carried out in which the effects of the order of the conditions were extracted from the error term. As can be seen in Table 1, the main effect of order of conditions was statistically significant. In addition, a significant interaction

TABLE 1
ANALYSIS OF VARIANCE FOR THE ASSEMBLY
EXPERIMENTS

Source	<i>df</i>	<i>SS</i>	<i>F</i>	<i>p</i>
Stereomonitors	1	105.14	—	
Order	1	707.64	4.52	<i>p</i> .05
VC × O	1	868.32	5.54	<i>p</i> .025
Error	44	6,896.30		
Total	47	8,577.40		

between viewing conditions and order of conditions (VC × O) occurred. Subsequent analysis showed that the effects of the order of conditions were primarily responsible for this interaction. However, even in this analysis using a more restricted error term, a statistically reliable difference between viewing conditions (stereotelevision versus single-camera viewing) did not occur.

It was assumed at the start of this experiment that the stereotelevision viewing would

aid the transport movements more than the grasp and placing movements, because the travel movements must be controlled in depth in the situation used. However, comparison of the durations of each component movement, as shown in Figure 2, failed to support this assumption. As the bar graph shows, times for stereotelevision viewing were slightly below those for single camera viewing for each component movement, but no one component was favored more than the others by the three-dimensional condition. Separate analyses of variance of the data for each of the four component movements revealed that none of the differences for viewing conditions were reliable.

EXPERIMENT II: ANGULAR DISPLACEMENT
OF STEREOTELEvised FEEDBACK
IN PERFORMANCE

This experiment's purposes were to explore the application of stereotelevision to sensory-feedback analysis of displaced perception and

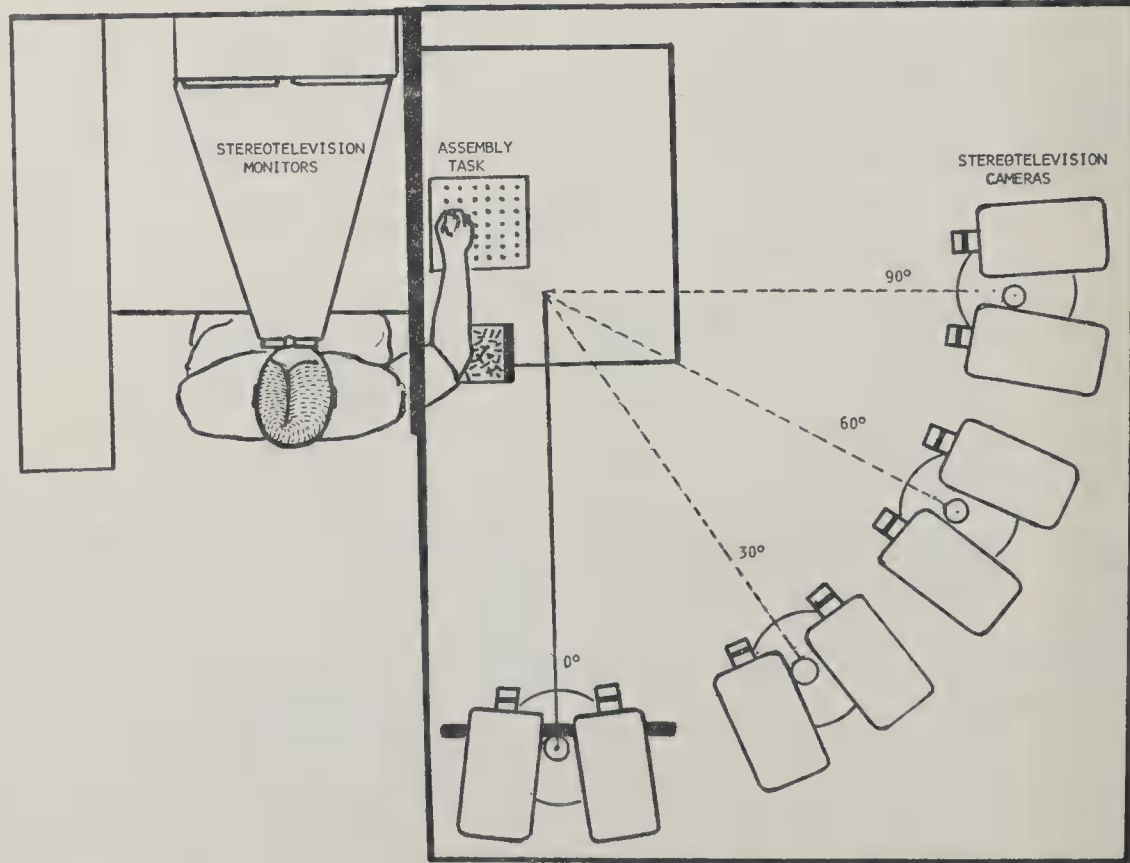


FIG. 3. The technique of angularly displacing the stereotelevised feedback of object assembly motions.

motion and to determine the effects of stereotelevised angular displacement upon object manipulation. The study is significant in indicating to what extent the control of the movements in object manipulation and assembly is dependent upon specific thresholds of relative angular displacement of motions and their visual feedback in depth. The techniques are of special significance because they represent the initial application of controlled methods of displaced three-dimensional visual-feedback analysis to the investigation of behavior organization.

Method

This experiment consists of angularly displacing the stereoscopic feedback of assembly motions in the horizontal plane of the performance field. The specific methods used are illustrated in Figure 3. Four angular displacements of the stereotelevised feedback of the assembly motions were used: 0, 30, 60, and 90 degrees. Two general conditions of visual feedback were employed with all four angular displacements: single-camera viewing, as described previously, and stereotelevised feedback. The line of regard of the cameras in the vertical plane was that normally occurring when *S* viewed the performance situation when in a sitting position at the work table. The general effect of the angular displacements of vision

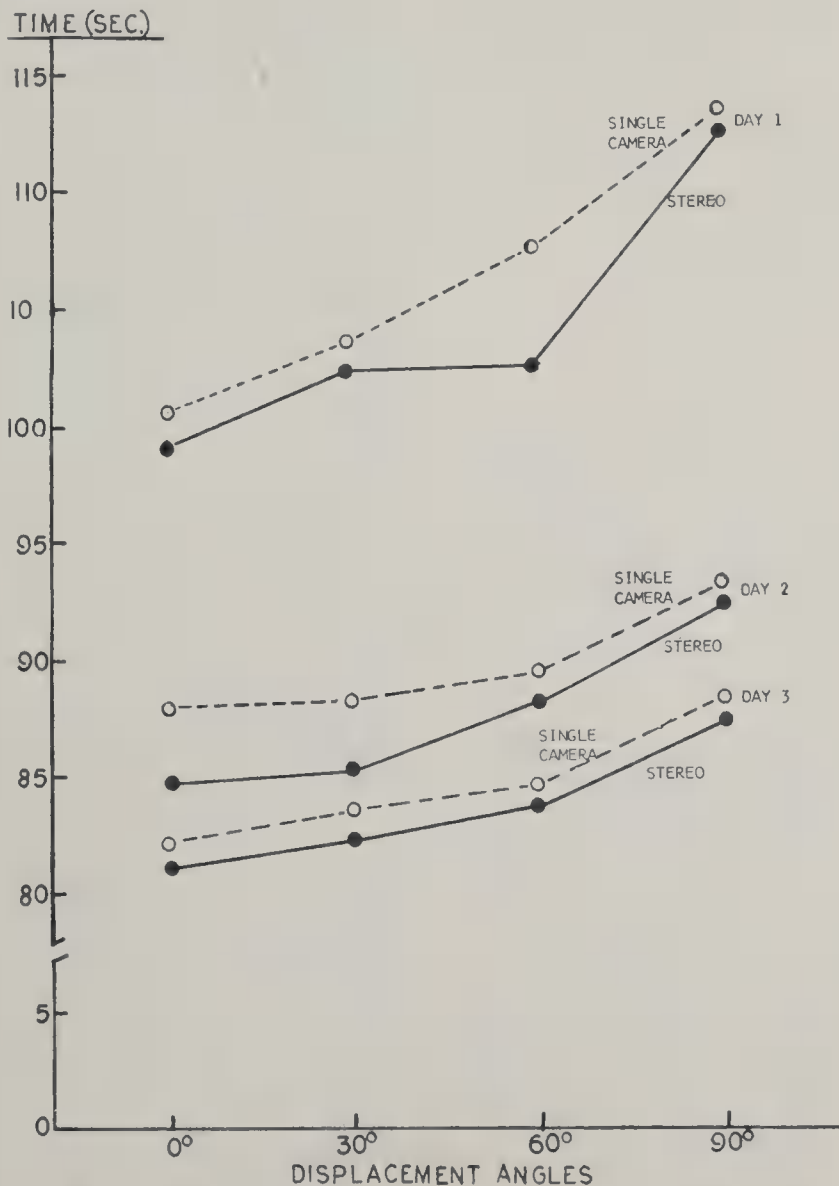


FIG. 4. Curves of motion time in object manipulation with single-camera viewing and stereotelevised feedback as a function of angular displacement of the visual effects of the motions. (The three sets of curves represent performance on the three successive days of practice.)

TABLE 2
ANALYSIS OF VARIANCE OF THE DATA ON THE ANGULAR
DISPLACEMENT STUDY

Source	df	SS	F	Error term
Viewing Conditions (VC)	1	507.28	1.60	6
Displacement Angles (DA)	3	6,093.26	26.70**	8
Subjects (S)	23	68,824.83	769.25**	(15)
Days of Practice (DP)	2	46,718.15	180.24**	10
VC × A	3	41.75	—	11
VC × S	23	7,288.57	—	13
VC × DP	2	35.38	—	13
DA × S	69	5,231.02	—	14
DA × DP	6	770.71	14.85**	14
S × DP	46	5,961.48	—	15
VC × DA × DP	69	8,935.09	—	15
VC × DA × DP	6	283.15	12.13**	15
VC × S × DP	46	6,360.37	—	15
DA × S × DP	138	1,194.05	—	15
VC × DA × S × DP	138	636.46	—	15
Total	575	158,498.46	—	15

** $p \leq .01$.

in the manner described is equivalent to removing the eyes of *S* from his head and placing them at the different angles of displacement mentioned. The experiment achieves for the first time the effective dislocation of binocular vision from the motor system of the body, and it analyzes in a preliminary way the effects of this dislocation.

The pin assembly task and electronic motion analysis setup previously described were used in this investigation. Twenty-four *Ss* each received eight trials of practice per day under the eight different conditions of viewing and angular displacement. Practice was continued for 3 days. Order of displacement conditions was randomized by subject in order to control for the effects of this order.

Results

The critical findings of the experiment are summarized graphically in Figure 4. This graph plots motion duration in seconds as a function of displacement conditions. The three sets of curves shown represent the 3 days of performance. The broken line represents single-camera viewing and the solid line is for stereotelevised feedback.

The analyses of variance of the data represented in Figure 4 are summarized in Table 1. The first feature of note in the graphs is that

performance with stereotelevised feedback is consistently superior to single-camera viewing at all displacement angles on all 3 days of practice. As indicated in Table 2, however, this main effect turns out to be statistically unreliable in terms of a 5% criterion level of significance. The *F* value for this main effect is based on the use of the first order interaction, Viewing Conditions × Subjects, as the error term. If, however, one resorts to pooling the higher order interactions (which may be properly done a priori) and uses this mean square as an error term to test the main effect of viewing conditions, then the difference between single-camera and stereotelevised feedback is found to be statistically significant. Thus, a trend exists in favor of the stereoscopic feedback condition.

As shown in Table 2, the main effect of Angles of Displacement was significant below the .01 level of chance. As the graph in Figure 3 illustrates clearly, time of performance consistently increases as the visual displacement is increased from 0 to 90 degrees. This variation occurred on all 3 days. In general, most *Ss* were able to obtain good fusion even with the severe displacement conditions, so that decline in performance at these conditions cannot be attributed to blurred vision.

The effect of practice is also statistically significant. The slopes of the curves in Figure 4 suggest and the statistical significance of the interaction Displacement Angle × Days observed in Table 1 indicates that the effect of displacement conditions upon performance was reduced with practice.

The different component movements of grasp, loaded transport, placing, and empty travel were all similarly affected by the two conditions of single-camera and stereoscopic viewing. That is, movement duration was consistently faster with the stereotelevised condition. However, none of these differences were statistically significant when they were tested against the first order interactions, Viewing Conditions × Subjects, as the error terms.

This study confirms in certain ways prior observations on the effects of angular displacement of the visual feedback of motion (Gould & Smith, 1962; Smith & Smith, 1962). These observations show that with

angular displacement of visual feedback of different motions a normal range of displacement is found that has no significant effect on movement time and accuracy, but that beyond this normal range a breakdown in performance occurs which reduces accuracy and increases the duration of movement. This breakdown angle has been observed to be much larger for assembly motions than for other more precise motions, when single-camera viewing is used. The results of this study generally confirm this finding with stereotelevised feedback. The results show that with displacement angles of 60 degrees and below Ss tend to adapt through practice to the displayed vision. However, with angles of displacement above this value the ability to adapt may be relatively slower and possibly never complete.

DISCUSSION

The present series of experiments was concerned with applying sensory-feedback methods of analysis to different performances with stereotelevised feedback. Such research has the main object of disclosing the dynamic real-time and real-space characteristics of behavior in controlling the visual feedback of movement with three-dimensional remote displays.

Although three-dimensional television consistently aids performance in the different component movements of an object assembly task over that found with single-camera viewing, this effect is not statistically reliable.

The effect of angular displacement of stereotelevised feedback of assembly motions is much the same as that found for displaced feedback with single-camera viewing. In both, a normal range of displacement is found in which motion is disturbed in only limited ways and practice is effective in eliminating the effects of the displacement. With angles of displacement of increased magnitude, however, movement organization is more severely disturbed and practice fails to eliminate these disturbances in the organization of motion. The breakdown displacement angle of object manipulation in the horizontal plane of the performance was found to be about 60 degrees.

Overall, the results of sensory-feedback studies are of importance in indicating the

nature and utility of stereotelevision as an experimental instrument for the study of the human factors problems of optical design in space science and in the industry of lethal materials. In these areas, the phenomena of space-displaced and delayed three-dimensional visual feedback of motion are of paramount significance for developing a systematic theory of machine design and for devising training simulation systems. Although it is possible to think of the problems of these fields as resolvable through the application of physical optics and conventional psychophysical methods of study, we have good reason to believe (Smith & Smith, 1962) that the behavioral problems of space-displaced and delayed vision can be solved only by the use of real-time and real-space analyses of the dynamic stimulus-feedback processes of performance and behavior. In all such human control systems, performance and motion are specialized products of the particular dynamic visual-feedback relationships that occur between the optical guidance system and the control motions performed. They are not direct functions of abstract psychophysical judgments and limits or the geometric properties of the optical system. Accordingly, psychophysiological research on the design of stereotelevision motion-control systems is a necessity if the design of such devices is to be optimized for both earth use or application to space technology.

CONCLUSIONS

1. Assembly motion with video feedback is consistently improved over monocular televised sensory-feedback control with stereotelevision, but the magnitude of the effects was not statistically reliable.

2. Angular displacement of the stereotelevised feedback of object assembly has much the same effect as similar angular displacement of single-chain viewing. The effects of such angular displacement of the visual feedback of motion provide direct evidence for the view that the direction, control, and learning of sensory motor response is based on the relative geometric displacement between movement and its sensory feedback.

3. Inasmuch as human visual behavior is specialized in terms of the dynamic movement

interactions between the optic and motor systems, specific psychophysiological sensory-feedback research is needed to explore the phenomena of stereotelevision in the design of behavioral control systems. The experiments reported illustrate significant methods of sensory-feedback analysis in this general field of optical design.

4. The present research is one step toward development of anthropomorphous optokinetic machines, which can be used both for mobile remote vision and for specialized delayed vision studies on the optic system itself.

REFERENCES

GOULD, J. D., & SMITH, K. U. Angular displacement of the visual feedback of motion. *Science*, 1962, 137, 619-620.

GOULD, J. D., & SMITH, K. U. Sensory-feedback analysis of stereotelevision pursuit tracking. *J. appl. Psychol.*, 1964, 48, 152-160.

MAURO, J. A. Three-dimensional color television system for remote-handling operation. In, *Human factors of remote handling in advanced systems: A symposium*. Wright-Patterson Air Force Base, Ohio: Aeronautical Systems Division, 1961. Pp. 103-168.

MENGLE, L. I. 3-dimensional TV system. *Radio TV News*, 1958, 60(4), 45, 128.

SMITH, K. U. *Delayed sensory feedback and behavior*. Philadelphia: Saunders, 1962.

SMITH, K. U. Sensory-feedback analysis in visual science: A new theoretical-experimental foundation for physiological optics. *Amer. J. Optom.*, 1963, 40, 365-417.

SMITH, K. U., & SMITH, W. M. *Perception and motion: An analysis of space-structured behavior*. Philadelphia: Saunders, 1962.

(Received August 28, 1963)

STEREOSCOPIC TELEVISION PURSUIT TRACKING

JOHN D. GOULD¹

*University of Wisconsin*²

This research was concerned with tracking a remote target moving in depth. A 3-D television system provided visual feedback, and direct and aided pursuit tracking systems were evaluated as a function of target speed. The stereoscopic display was shown to be generally satisfactory for remote-control operations, although some fatigue or "eyestrain" was reported, probably due to the optics of the system. Contrary to previous tracking studies on nondepth courses, it was shown that on a depth course direct tracking is consistently superior to aided tracking at the 3 target speeds used. Amplitude of error analysis provided answers concerning what Ss do when not on target. Tracking behavior was interpreted in terms of the sensory-feedback mechanisms governing the control of motor patterns.

This research has two main purposes: to assess the usefulness of a recently developed binocular-separation stereoscopic television system for remote operation and visual research, and to compare aided and direct pursuit tracking on a depth course as a function of target speed. Accordingly, subjects (Ss) used either a direct or aided tracking system in tracking a remote target moving in depth at one of three target speeds while viewing their performance on the stereoscopic visual display.

Several types of three-dimensional displays have been utilized to present feedback of remote-control operations, including three separate dials; a dual-beam cathode-ray tube (CRT) (Morrill & Davies, 1961); and two independent CRTs where the operator fuses the information presented on each tube by means of prisms. In addition, mechanical-replica displays have also been suggested, where the operator views a miniature replica, varying in three-dimensions, of the object under remote control (Bauerschmidt & Bescoe, 1962). Applications of the kinetic depth effect (Fried, 1960) and also "volumetric" displays (Bassett & Stone, 1962),

where a fixed CRT projects onto a screen rotating inside a clear hollow solid, have been studied for their three-dimensional value.

Pictorial stereoscopic systems have essentially centered around the use of television. These date back to 1928 and include the use of separate projection systems transmitted to the operator *sequentially* on a single display; *simultaneous* transmission by two television cameras to a single television monitor; two monitors at 90 degree angles, viewed through a half-silvered mirror where the operator fuses the two disparate pictures using polarized light (Johnston, Hermanson, & Hull, 1950); and the use of color television (Mauro, 1961). Projection television methods and miniature television "eyes" may also be used in three-dimensional pictorial displays. The present research utilizes a binocular-separation stereoscopic television system where each of two television cameras converge symmetrically upon a target and feed their signals into separate, adjacent monitors. The operator then fuses the disparate images by means of prisms. This three-dimensional system has already been subjected to considerable human factor analysis (Gould & Smith, 1964).

Previous tracking studies, wherein an operator attempts to align an indicator or cursor with a continuously moving target, have generally shown that direct tracking is superior to aided tracking³ on complex and high speed

¹ This research, submitted as a doctoral dissertation at the University of Wisconsin, was supported in part with the National Institute of Mental Health and National Science Foundation grants of Karl U. Smith, and it was conducted while the author was a National Institutes of Health research fellow. The author wishes to express his gratitude to Karl U. Smith for his guidance on this project.

² Now at the Thomas J. Watson Research Center, IBM, Yorktown Heights, New York.

³ In aided tracking the operator directly positions, by means of a hand control, the cursor or indicator, and in addition, simultaneously inserts an automated rate of cursor movement. The automated rate de-

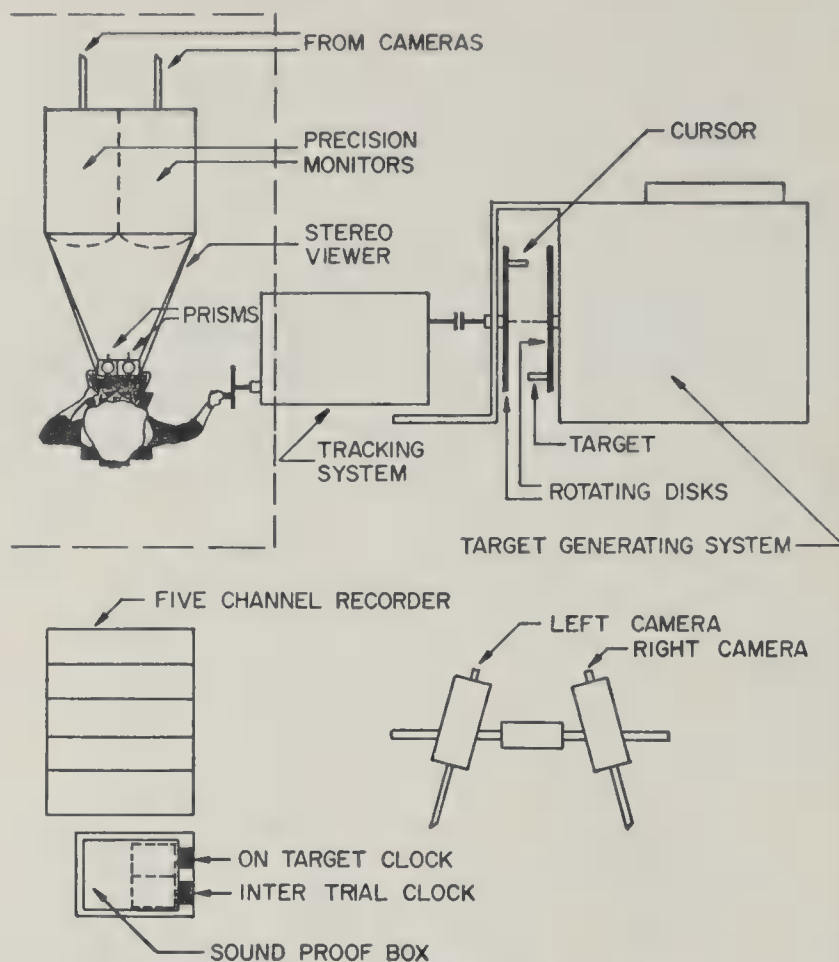


FIG. 1. Overhead view of the experimental layout. (The operator is seated in an enclosed cubicle while viewing his remote performance via stereoscopic televised feedback.)

courses, whereas the opposite relation holds true for relatively simple, slow-moving targets (Chernikoff & Taylor, 1957; Lincoln & Smith, 1952). The present research evaluates this generality for binocular tracking of a target moving in depth, a problem that has received little experimental attention although the visual feedback differs markedly from that of frontal plane tracking.

METHOD

Apparatus. An overhead view of the experimental layout of the stereoscopic television system and

depends upon the aided time constant which is defined for a given displacement of the hand control as the ratio of direct cursor displacement to the automated velocity component. For example, an aided time constant of .25 second means that a certain hand-wheel displacement which directly moves the cursor 15 degrees simultaneously generates a cursor velocity of 60 degrees per second. The time constant gives the time necessary to automatically drive the cursor through the arc of initial displacement.

depth tracking system is shown in Figure 1, where the apparatus used may be considered under four headings.

Depth Target Generating System. The course generating system used to move a target in depth was modified from Lincoln and Smith (1952). The target, a 4-inch white rod, .5 inch in diameter, was mounted perpendicular to a vertical black disk, 8 inches from the center of rotation of the disk. The center of the target disk, the edge of which faced the front of the display, was mounted directly to the shaft of a ball-and-disk integrator. A variable speed synchronous motor drove the disk of the integrator, the speed of this motor being switched for the three target speeds used. A 1-rpm synchronous motor drove a cam which continuously changed the position of the ball bearings on the integrator disk thus causing the target to move in a circular fashion, toward and away from the front of the display. The shape of the cam was such that nine reversals of direction and continuous changes in velocity of the target occurred. The average speed of the three target courses was 5.67, 11.33, and 22.67 degrees per second, all three being based upon the same cam drive. These speeds were selected such

that the lowest would be considered "slow" relative to previous studies and the highest as "fast."

Tracking Control System. In the direct tracking system *Ss* remotely controlled the cursor by means of two pulley systems. Preliminary investigations indicated a 3-1 handcrank-cursor ratio as most optimal for the depth task. The cursor itself was identical to the target, occupying 3.58 degrees in its 360 degree orbit, and was also mounted perpendicular to a similar black disk parallel to that of the target. Both the target and cursor rotated independently in similar orbits, and their tips were separated by 1 centimeter when aligned. The combination of direct positioning and automated rate control that occurs in aided tracking was accomplished by means of a ball-and-disk integrator, a mechanical differential, and a synchronous motor, the voltage being kept at a fixed level. Rotation of the *S's* handcrank directly positioned the cursor and simultaneously changed the location of the ball bearings in relation to the center of the integrator disk, the disk being driven at a constant speed by the synchronous motor. This change in location of the ball bearings was reflected in a change in direction and/or velocity of the output shaft of the integrator. Both the direct positioning effect and the automated velocity effect served as the input to the mechanical differential through which the cursor was controlled.

Display and Stereoscopic Television System.

Figure 1 shows the position of the two RCA vidicon cameras (model TK-201), each of which fed their independent signals to corresponding adjacent Conrac (model CN A8/C) 8-inch monitors located in front of the operator. The interaxial lens separation of the cameras was 12.5 inches and the midline distance from the center of the target-cursor depth display to the cameras was 90 inches. Mounted on the rasters of the two adjacent monitors was a black tapered viewing hood, 25 inches deep, the narrow end containing a set of adjustable rotating prisms by which the operator fused the two similar but disparate televised images into a three-dimensional scene.

Each white rod appeared on the television screens to be 3.75 to 4.25 millimeters wide depending on its distance from the front of the depth tunnel. The distance between them when aligned appeared between 2 and 2.5 millimeters. Previous studies (Gould & Smith, 1964) have shown that the visual acuity functions of the stereosystem are about half that of normal viewing and that the stereoscopic depth threshold is considerably less than in normal viewing, the average absolute error depending upon the interaxial separation of the cameras.

Error Recording System. Two measures of performance were used in this experiment: time on target and amplitude of error distributions. The time-on-target clock, triggered by a photocell circuit in the tips of the target and cursor, was located,

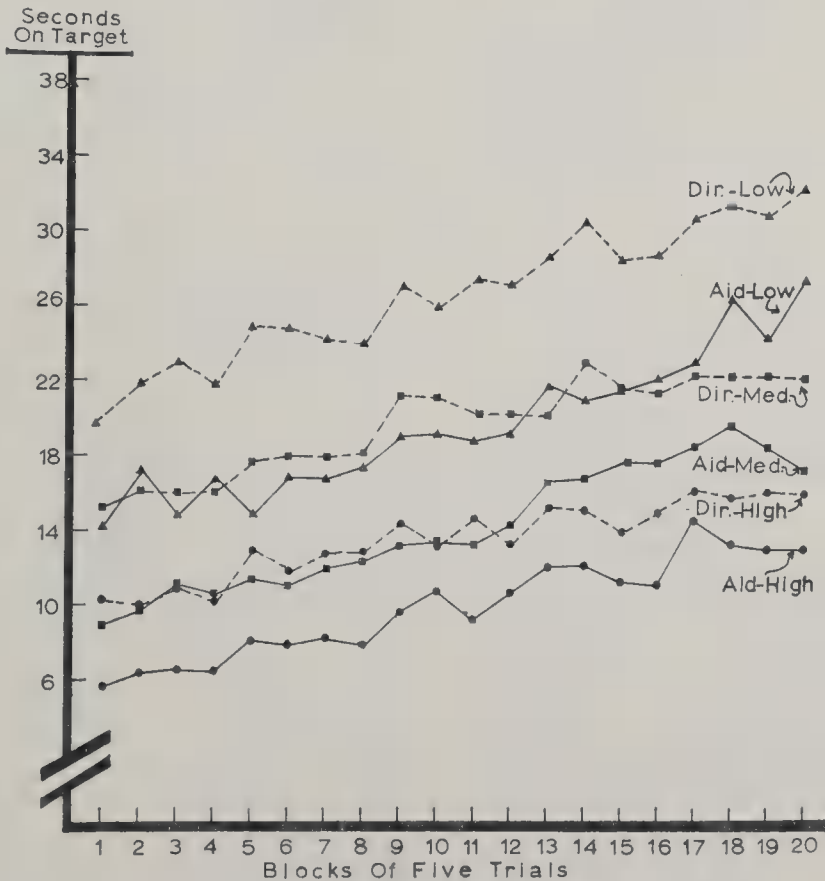


FIG. 2. Time-on-target scores for the three target speeds and two tracking systems.

together with all relays, in a soundproof box to prevent auditory feedback of tracking performance. "On Target" was adjusted to record all alignments of ± 1.2 degrees between the centers of the target and cursor. Graphic tracking error records were obtained by mounting the body of a continuous potentiometer directly to the center of the target disk while the shaft of the pot was connected to the center of the cursor disk. Thus both the body and shaft of the pot rotated independently, and any given difference between the target and cursor caused the same voltage signal at the polygraph error recorder (Offner Type R Dynograph) regardless of the actual position on the depth course.

Procedure. The repeated measures group design used in this experiment consisted of 48 volunteer undergraduate Ss divided equally into six experimental groups corresponding to the three target speeds and two tracking systems. Each S tracked 20 1-minute trials on five consecutive days. Within a day each trial was separated by 30 seconds, this interval being automatically timed. At the beginning of each trial the target and cursor were aligned, although their absolute position on the depth course was varied throughout. Each S was thoroughly familiarized with the equipment and shown how to fuse the stereoscopic images. Particular care was also taken to insure that Ss using the aided system understood its operation.

Scoring. The 4,800 graphic error records were scored manually by six paid undergraduates, each of whom scored about an equal number of records from all six experimental conditions. The amplitude of error on each 1-minute trial was sampled at 2-second intervals, making a total of 144,000 error samples. Besides the On-Target samples falling within ± 1.2

degrees of target-cursor alignment, errors were classified into 12 additional amplitude categories: $\pm 1.2-3$, $\pm 3-6$, $\pm 6-9$, $\pm 9-12$, $\pm 12-15$, $\pm 15-18$, $\pm 18-30$, $\pm 30-60$, $\pm 60-90$, $\pm 90-120$, $\pm 120-150$, and $\pm 150-180$ degrees. If an error fell on a boundary between two categories it was scored as being in the larger direction.

RESULTS

Effects of Aided and Direct Tracking in Depth. A main result shown in the time-on-target curves of Figure 2, where each data point is based upon 40 scores, is that learning occurred in all six experimental conditions. While approximately the same amount of learning occurred on all three target speeds, significantly more practice effects were found with the aided tracking system. Yet direct tracking of a depth target remained superior to aided tracking throughout practice. In fact, performance in the Aided-Medium condition did not consistently exceed that of the Direct-High condition until 60 practice trials, and 75 practice trials were necessary before the Aided-Low condition exceeded the Direct-Medium condition.

Overall, time-on-target scores showed that direct tracking was superior to aided tracking on a depth course. This generalization was true at each target speed also, where independent *t* tests showed direct tracking to be

TABLE 1
SUMMARY OF ANALYSIS OF VARIANCE

Source	df	MS	F Test
Tracking System	1	36,775.53	19.32**
Target Speed	2	52,796.35	27.73**
System \times Speed	2	1,286.13	— —
Ss/Conditions	42	1,903.79	82.67**
Trials	(99)	434.0	18.85**
Between Days	4	10,144.68	94.34**
Within Days	19	25.68	1.36
Between Days \times Within Days	76	25.08	1.28
Trials \times Systems	99	30.75	1.33*
Trials \times Speed	198	25.48	1.11
Trials \times Systems \times Speed	198	26.03	1.13
Ss \times Trials/Conditions	(4158)	23.03	
Ss \times Between Days/Conditions	168	107.53	
Ss \times Within Days/Conditions	798	18.92	
Ss \times Between Days \times Within Days/Conditions	3192	19.61	
	4799		

* $p < .05$.
** $p \leq .01$.

significantly better beyond the .01 level of confidence. Thus, the generalization based upon nondepth courses where aided tracking was found to be superior to direct tracking at low frequencies of target movement did not hold for a target moving in depth. The degree of superiority of direct tracking was approximately the same at each target speed, there being no interaction between target speed and tracking system.

The analysis of variance of time-on-target scores, summarized in Table 1, verifies the above statements. As may be seen, significant differences in performance occurred because of the tracking system used and also because of target speed. While learning occurred in all six tracking conditions the statistical analysis showed the interaction between Tracking Systems and Trials was significant, where relatively more learning occurred on the aided system. However, Target Speed did not interact with either Tracking System or Trials. The curves of Figure 2 were further broken down into the amount of learning within each day of practice (20 trials) and the change in performance between days to estimate the "fatigue" involved using the three-dimensional feedback system. The analysis of Table 1, using the appropriate error terms, shows that no significant improvement in tracking occurred within each day indicating that Ss performed at about the same level throughout any one practice session. However, the significance of the Between-Days variable shows that the learning functions of Figure 2 were due to increases in the level of performance between days of practice.

Amplitude of Error as a Function of Tracking System and Target Speed. Time-on-target scores give no indication of the degree and type of error during time off target. The significance of analyzing the amplitude of error during tracking lies in discovering the differences in types of error as a function of experimental variables. In Figure 3 the log number of samples falling into the On-Target category as well as the 12 error amplitude categories is plotted on the ordinate for each of the six depth tracking conditions. The curves are quite orderly in showing the differential effects of target speed and tracking system on the error functions of each experimental

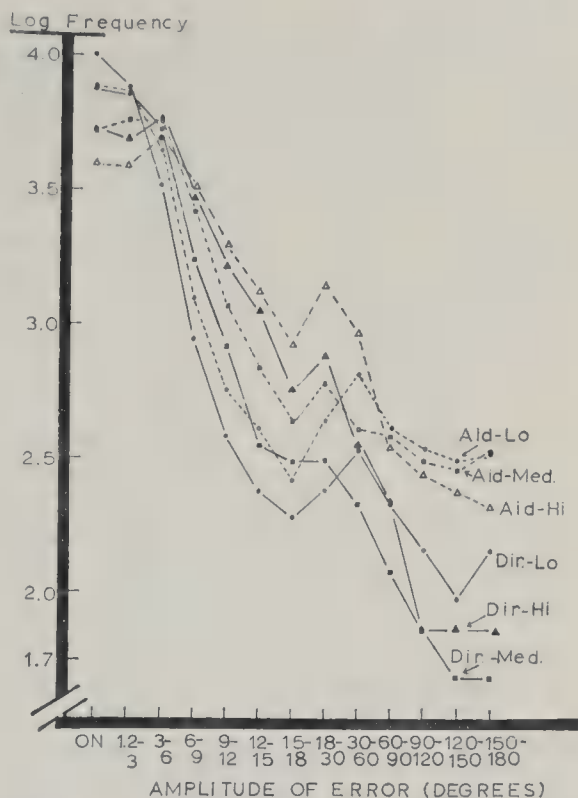


FIG. 3. Overall amplitudes of error distribution for the six experimental conditions.

condition. Target-cursor alignments between ± 3 degrees were a joint function of target speed and tracking system, where more of these alignments occurred with the direct system at each target speed and the frequency of these alignments decreased with increases in target speed. Intermediate errors of ± 6 – ± 18 degrees showed just the inverse effect, while gross errors larger than 30 degrees were specifically determined by the tracking system used. Chi-square analysis showed the statistical significance of this interaction between tracking systems, target speed, and amplitude of error. This interaction is of much importance for it demonstrates that movement organization in tracking is not invariant nor independent of perceptual-operational and instrumental feedback effects. Rather, tracking errors are dynamic and interacting, depending specifically upon the feedback effects involved.

Change in Amplitude of Error with Practice. The rate of change of small errors with practice in depth tracking was found to be a function of the relative size of these small errors. That is, by plotting percentage time on

TABLE 2
PERCENTAGE INCREASE IN PERFORMANCE PER BLOCK OF TRIALS

Error criterion	Tracking condition					
	Aided-Low	Aided-Medium	Aided-High	Direct-Low	Direct-Medium	Direct-High
±1.2 degrees	1.10	1.46	1.33	1.07	1.42	1.27
±3 degrees	2.74	2.51	2.60	1.88	2.22	2.21
±6 degrees	1.94	2.80	3.68	1.82	1.73	2.07
±9 degrees	1.84	2.47	3.68	1.44	1.34	1.88
±12 degrees	1.70	2.11	3.54	1.34	1.09	1.48
±15 degrees	1.52	1.84	3.26	1.32	.94	1.27

target against trials for each of the error categories it was determined that in both aided and direct tracking the differences between the 6, 9, 12, and 15 degree functions became larger as target speed increased. There was a general tendency with the direct tracking system for small errors less than ± 3 degrees to increase with practice while larger errors decreased; with the aided system, however, errors up to ± 6 degrees increased with practice.

These results may also be considered in terms of the method used to assess tracking performance, where each error category is a

possible cut off or criterion measure for time-on-target scores. A quantitative index of these cumulative plots of tracking accuracy over time, determined by the method of successive differences, is shown in Table 2. The value of each cell represents the percentage increase in tracking performance per block of trials. This measure is more meaningful than an overall slope coefficient, which is dependent upon the arbitrary distance between the blocks of trials on the abscissa of any graph. The rate of improvement, trials to asymptote and naturally the overall performance level in tracking was a function of the criterion used.

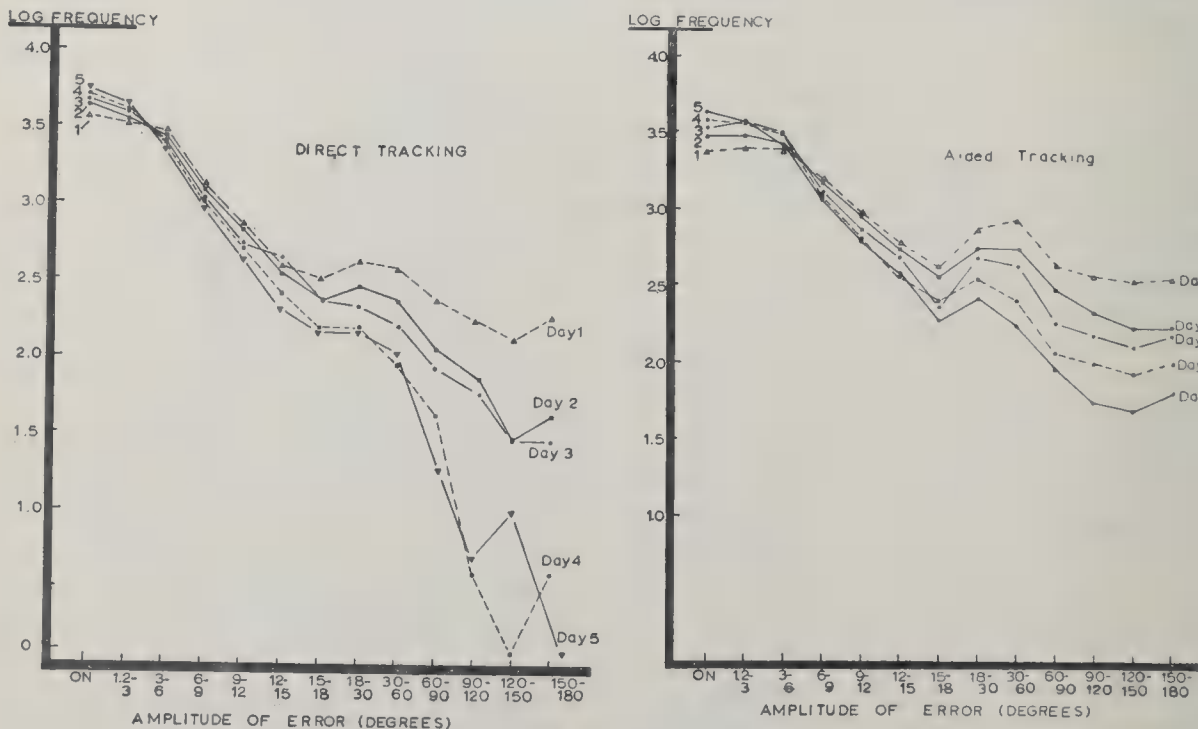


FIG. 4. Change in error amplitude for each tracking system as a function of practice.

With both aided and direct tracking systems little difference in performance occurred as a function of target speed at the end of practice if ± 15 degree deviations were used as the criterion measure. Likewise, the fact that errors between ± 3 – 6 degrees increased with practice in aided tracking and decreased in direct tracking (Figure 4) would enhance the former condition if ± 6 degrees was used as the criterion. A more lax criterion thus reduced the effects of both target speed and tracking systems. These findings show that learning curves, as well as the differences between them, are dependent to a great extent upon the criterion used to assess performance (Bahrick, Fitts, & Briggs, 1957).

The two graphs in Figure 4 show the change in large as well as small error amplitude with practice for the direct and aided systems. The parameter of the curves is days of practice, each day being based upon 20 trials; the log number of samples falling into On-Target and the 12 error categories is plotted on the ordinate. The main result to note is that large errors over 30 degrees decreased much more rapidly with practice using a direct tracking system than an aided system, even though there were significantly more of these errors to be reduced with the aided system from the outset.

DISCUSSION

Stereoscopic Television System. The results of this research showed that Ss were able to maintain fusion of the two disparate television images while tracking a depth target. The absolute tracking errors gave an indication of the accuracy of the stereoscopic visual feedback provided. After 4 days of practice with direct tracking Ss were able to keep the cursor within ± 3 degrees of the target, which corresponded to at least some overlap of the target and cursor, over 50% of each trial on the high-speed course, one which is quite complex and seemingly difficult. With the same amount of practice, Ss maintained the same criterion over 70% of the time on the medium-speed course and over 74% of the time on the low-speed course. Unfortunately, the error amplitude values could not be translated into subtended visual angle, as these angular values changed for any given absolute

deviation as a function of the relative position of the target and cursor on the depth course.

It is felt that the results obtained in this study together with those of previous research (Gould & Smith, 1964) using this stereoscopic system indicate its value for precise remote-control work. While no comparison was made with normal viewing, the assumption being that direct viewing is superior, previous studies (Gould & Smith, 1964) have consistently shown the superiority of this stereoscopic television system over a single camera-monitor system for remote-control work. In addition, this stereosystem provides for the regulation and control of binocular visual feedback in scientific behavioral studies. It is now possible to systematically and independently control, displace, distort, and regulate the continuous spatial and temporal properties of visual input to each eye.

A weakness of the present three-dimensional system was reflected in the fact that Ss showed little improvement in performance throughout any one day of practice, a result that was not entirely anticipated. It appears that Ss were learning during each 20-trial session, since each subsequent day's performance was superior to the previous day's, but this learning was somehow suppressed within a day due to the "eyestrain" and fatigue resulting from the stereoscopic system. While it is felt that the small distortions introduced by the wide interaxial separation of the television cameras and monitors, the use of prisms, and the resolution of the television system contributed to this eyestrain, it is believed that the disruption of the accommodation-convergence mechanism, due to the continual changing of the latter with fixation while accommodation remained relatively constant, is the prime factor. There is the possibility that the stereoimages were not optimally placed or framed also. A certain amount of depth "confusions" occurred, i.e. where both target and cursor appeared in the same horizontal plane but one was at the front of the display and the other at the rear, which corresponded to some of the large amplitude errors over 45 degrees. However, these depth confusions rapidly dropped out with practice (Figure 4), at least with the direct tracking system. The Ss generally expressed much

satisfaction with the accuracy of the feedback provided. Only two Ss had to be dropped from the experiment because of inability to maintain fusion, one because she could see with one eye only.

Tracking Behavior. The results of this study specifically showed that, using time-on-target scores, a direct tracking system was superior to an aided system throughout practice in depth tracking, and that this superiority was about the same at all target speeds. That the slow-speed course was actually slow relative to previous studies can be shown by the fact that the target moved a total of only 340 degrees in a 1-minute trial. It was further shown that relatively more learning occurred with aided tracking than direct tracking; speed of target movement did not interact with practice or tracking systems, a result that was not predicted because of previous results in nondepth tracking (Chernikoff & Taylor, 1957; Lincoln & Smith, 1952) where the general conclusion is that direct tracking becomes increasingly superior as speed increases. It was also shown that amplitude of target-cursor error interacts with both tracking systems and target speeds.

A possible criticism of these findings, one which can be leveled at any tracking study, is the choice of the handcrank-cursor ratio and aided time constant, even though pre-experimental investigations determined those used as optimal. Since the time constant (2.00 seconds) was higher than the usual .5 seconds, the absence of an interaction between target speed and tracking system could be explained on the basis of Chernikoff and Taylor's (1957) findings as due to a compensating effect, since they showed an upward shift in the optimal time constant with increases in target frequency, and this would tend to enhance aided tracking at high speeds relative to low speeds.

However, it is preferred here to interpret this result somewhat differently. Besides the fact that Pearl, Simon and Smith (1955), in contrast to Chernikoff and Taylor (1957), indicated a *downward* shift in the optimal time constant as target speed increased, previous studies have not evaluated pursuit tracking in depth, a perceptual-control problem that is by definition quite different

because of the spatial feedback involved, as witnessed by the front-rear depth confusions that occurred at all target speeds.

It is believed here, in conjunction with a previous body of experimental results (Smith, Gould & Wargo, 1963; Smith & Smith, 1962), that all human tracking behavior is organized in terms of sensory-feedback mechanisms which regulate visual-motor organization and the component torso, limb, hand, and finger movements. These manual movements are made up of high-speed tremor movements, precise positioning movements of the hand and fingers, intermediate movements which are a combination of positioning and rate movements, and gross rate movements of the hand and arm. The sensory feedback effects of tracking are dependent upon the component movements involved (reactive feedback); the perceptual dynamics of the tracking system (instrumental feedback), such as aided versus direct, display gain, etc.; and the action of the controlled cursor or indicator relative to the target or target path (operational feedback) which includes the speed and complexity of the target path as well as the spatial and temporal characteristics of the visual display used in remote tracking.

By varying instrumental and operational feedback effects, such as has been done in this tracking task, certain inferences may be made concerning movement organization and its dependence upon sensory feedback. In terms of the categories of errors used in the present research, it is proposed that large tremor and very fine positioning errors may well correspond to ± 1.2 –3 degrees of target cursor deviation. The larger proportion of positioning errors probably cause target-cursor deviations of ± 3 –18 degrees. Intermediate movement errors are described by ± 18 –30 degrees of error, and large errors in rate control movements cause the deviations between ± 30 –180 degrees.

The results of this study show that very small positioning and tremor errors in depth tracking were determined mainly by the operational feedback of target speed. However, more of these errors occurred at each target speed for the direct system than the aided system. Apparently, the aided system was either successful in filtering out these

small errors as aided systems are generally designed to do or else such a system reduces the possibility of spending as much time within ± 3 degrees of the target. That the latter is the case can be seen in Figure 4 where small errors increased with practice on the aided system. Errors in positioning movements (3–18 degrees) were also a joint function of the operational feedback of target speed and the instrumental feedback of tracking system; however, these positioning errors showed just the inverse function of that found for the errors of less than ± 3 degrees. Differences in rate control movement errors were determined specifically by the differential instrumental feedback provided by the tracking systems. A main difference between aided and direct tracking is the visual delay inherent in the former system. It has been maintained (Smith, 1962) that this delay is responsible for the inferior performance of the latter system since the operator must learn to predict, at least to some extent, the cursor movement as well as the target movement. If these inferences, based upon experimental data, are correct then no unique pattern of motion exists in the case of tracking tasks. Rather tracking, as well as other complex motor activity, is dependent upon the sensory feedback consequences of behavior.

REFERENCES

- BAHRICK, H. P., FITTS, P. M., & BRIGGS, G. E. Learning curves—facts or artifacts. *Psychol. Bull.*, 1957, **54**, 256–268.
- BASSETT, R. C., & STONE, J. T. Concepts and requirements for volumetric 3-dimensional displays. *International Congress of Human Factors in Electronics*, Institute of Radio Engineers, 1962.
- BAUERSCHMIDT, D. K., & BESCOE, R. O. Human engineering criteria for manned space flight: Minimum manual systems. *USAF AMRL tech. docum. Rep.*, 1962, No. 62–87.
- CHEERNIKOFF, R., & TAYLOR, F. V. Effects of course frequency and aided time constant on pursuit and compensatory tracking. *J. exp. Psychol.*, 1957, **53**, 285–292.
- FRIED, C. Studies on the kinetic depth effect as a means for presenting three-dimensional information: I. Methodology and selection of forms for study. *USA Ordn. Hum. Engng. Lab. tech. Memo.*, 1960, No. 2-60. (abstract)
- GOULD, J. D., & SMITH, K. U. Bisensory tracking via stereoscopic television. *J. appl. Psychol.*, 1964, in press.
- JOHNSTON, H. R., HERMANSON, C. A., & HULL, H. L. Stereotelevision in remote control. *Elect. Engng.*, 1950, **69**, 1058–1062.
- LINCOLN, R. S., & SMITH, K. U. Systematic analysis of factors determining accuracy in visual tracking. *Science*, 1952, **116**, 1–5.
- MAURO, J. A. Three-dimensional color television system for remote-handling operation. In, *Human factors of remote handling in advanced systems: A symposium*. Wright-Patterson Air Force Base, Ohio: Aeronautical Systems Division, 1961. Pp. 103–168.
- MORRILL, C. S. & DAVIES, B. L. Target tracking and acquisition in three dimensions using a two-dimensional display surface. *J. appl. Psychol.*, 1961, **45**, 214–221.
- PEARL, B., SIMON, R., & SMITH, K. U. Visual tracking: IV. Interrelations of target speed and aided-tracking ratio in defining tracking accuracy. *J. appl. Psychol.*, 1955, **39**, 209–214.
- SMITH, K. U. *Delayed sensory feedback and behavior*. Philadelphia: Saunders, 1962.
- SMITH, K. U., GOULD, J. D., & WARGO, L. Sensory feedback analysis of behavior. A new theoretical-experimental foundation of physiological optics. *Amer. J. Optom.*, **40**, 1963, 365–417.
- SMITH, K. U. & SMITH, W. M. *Perception and motion*. Philadelphia: Saunders, 1962.

(Received September 16, 1963)

VALIDATION OF THE MINNESOTA VOCATIONAL INTEREST INVENTORY FOR VOCATIONAL HIGH SCHOOL BOYS¹

W. LESLIE BARNETTE, JR., AND JOHN N. McCALL

State University of New York at Buffalo

The concurrent validity of the MVII was investigated with over 1000 vocational high school boys in Grades 9 and 12 in Buffalo, New York, schools. Scores of boys in particular trade curricula were checked against relevant MVII scales. At Grade 12, the food, electrical, and printing trade choices were well predicted; students in building trade, machinist, and mechanical programs were not well spotted. Similar results, but less encouraging, were found for the Grade-9 sample. With 1 student sample only (electrical), aptitude test data were unrelated to MVII scores. Students with "high" academic or shop school averages earned higher MVII criterion scale scores than did others.

This is the first of a proposed two-stage study of the interests of vocational high school students. The primary goal is to see if the Minnesota Vocational Interest Inventory (MVII) holds promise for the selection and counseling of students in vocational high schools. It is hoped these studies will help make clear the part which interests have in choosing and adjusting to these training programs and later vocations.

The MVII was standardized on adult workers; its value for high school boys remains unknown. This first study investigates the agreement between MVII scores and present enrollment in different vocational courses. The second study is planned to investigate the MVII's capacity to predict senior year status for freshmen tested earlier.

Several factors may decrease the practical validity of the MVII for high school boys. They may lack sufficiently mature or realistic interests, as measured by tests; outside administrative or socioeconomic pressures may determine the choice of given training programs; parental guidance may play a dominant role. It seems worthwhile to study the MVII's usefulness in an ongoing, realistic

program of vocational education, especially because such outside factors are likely to persist.

DESCRIPTION OF THE MVII

Full details of the construction and validation of the MVII are to be found in Clark's recent volume (1961). The inventory was developed to serve essentially as an instrument to measure interests in the skilled-trade area. At the outset, however, the possibility of development of other types of scoring keys was recognized (as that for retail salesclerk). The inventory consists of 570 items arranged in 190 triads. The forced-choice format requires the individual to select from each triad the one activity he most and least likes. Campbell and Sorenson (1963) have shown that this inventory is essentially free of any response bias caused by item location within the triad. Administration time varies from 45 minutes to 1½ hours per client.

Separate scoring keys are available for 20 occupations, mostly of the skilled-trade type: baker, carpenter, electrician, IBM operator, machinist, mechanic, milk wagon driver, painter, plasterer, plumber, pressman, printer, sheet metal worker, stock clerk, truck driver, warehouseman, retail salesclerk, food service manager, hospital attendant, radio-TV repairman. Aside from the general information that the early skilled-trade groups are from the St. Paul and Minneapolis area, possibly introducing a geographical bias into the results, little

¹The research reported here was supported through the Cooperative Research Program of the Office of Education, United States Department of Health, Education, and Welfare as Cooperative Research Project Number 1350. The junior author presented a brief report of this research at the American Psychological Association meetings in Philadelphia, Fall 1963.

is known concerning the specific composition of these criterion groups. The original work was done with eight AFL union groups and with the cooperation of both the union officials and employers. Inspection of the titles of some of the other scoring keys will clearly indicate that other, nonunion, nonskilled trade groups have been utilized. Presumably the composition of these will be completely described by Clark when the first Manual for the MVII becomes available (to be published by the Psychological Corporation).

The immediate relevance of many of these scoring keys to counseling needs of vocational high school boys (as well as for the guidance people in these high schools) is readily apparent. It has long been thought that differentiation of interests at that level—at least with current inventories—was impossible. Motto (1959), for example, using Kuder scores to predict vocational school success, found that these students produce essentially flat profiles and that none of the scales could be used to differentiate successful from unsuccessful students. If, with this more specially designed interest inventory, it could be shown that these adult criterion groups work at the vocational high school level, this would constitute an enormous help to all guidance people. Furthermore, as Wolfbein (1961) has shown, the number of skilled workers for the United States as a whole is expected to increase. The vocational high schools are the major suppliers of these workers. About three of every four craft workers are year-round full-time workers. The biggest contribution to present unemployment figures is the high vocational school dropout rate. Such facts point to even greater urgency for improvements in the counseling of vocational high school boys.

STUDENT SAMPLE AND DATA TREATMENT

The research design called for testing over 1,000 boys in the ninth and twelfth grades of four Buffalo vocational high schools. Major training programs in these schools were building, electrical, food service, machinist, mechanical, and printing trades. For most of these programs there were matching MVII scoring keys available. In addition to all 20 MVII scores for each boy, school records

provided data regarding general intelligence level (typically, Lorge-Thorndike IQs from the eighth-grade level) and scores on five of the DAT subtests. Average grades for academic and shop courses were also computed both for the ninth- and twelfth-grade samples.

Usable MVII scores were obtained for 1,114 boys at Grades 9 and 12. There are slightly more of the former than the latter (58% versus 42%). Almost 10% of the sample were Negroes which also allowed the authors to check for possible ethnic differences. All the boys were urban residents of Buffalo. By the boys' own description, most fathers were employed in skilled and semiskilled occupations; over one-third of the boys indicated the mothers worked outside the home. Grade-9 boys were tested in January 1962; twelfth graders in the Spring of 1962, just prior to graduation.

RESULTS AND DISCUSSION

Comparisons of mean MVII scores for Grade 9 versus Grade 12 and for Negro versus white produced few significant differences. The Grade-9 boys and the Negroes showed somewhat higher clerical and personal service interests than do either the Grade-12 boys or whites. Two-thirds of the Negro sample, however, are ninth graders. The differences in MVII scores are greater between the two grade levels than between the two ethnic groups.

Significant differences were obtained between ninth and twelfth graders on the aptitude and achievement measures but not for the general intelligence test. Scholastic aptitude test results suggest the more intellectually able students leave the vocational high schools; the spatial and mechanical test results suggest the more technically skilled students remain. The Negro subjects showed lower average scores than whites on all the ability measures save one (the DAT Abstract Reasoning), a test frequently regarded as lacking "academic" content.

Profiles of average MVII scores were determined separately for each trade at both grade levels. The means for the Grade-12 sample are given in Table 1 and have been transformed to correspond with Clark's civilian scales. Interpretation of these average

TABLE 1

PROFILES OF AVERAGE MINNESOTA VOCATIONAL INTEREST INVENTORY SCORES (ADULT NORMS)
FOR TWELFTH GRADE BUFFALO VOCATIONAL SCHOOL TRADES

Trade	N	Scale Baker	Carpenter	Electrician	IBM operator	Machinist	Mechanic	Milk wagon driver	Painter	Plasterer	Plumber	Pressman	Printer	Sheet metal worker	Stock clerk	Truck driver	Warehouseman	Retail sales clerk	Food service manager	Hospital attendant	Radio/TV repairman
Bricklaying	7	47	44	34	49	29	35	42	43	49	40	36	40	41	48	52	50	43	44	45	34
Carpentry	19	51	45	37	45	29	28	46	49	40	31	36	40	38	50	44	49	46	43	43	35
Plastering	6	47	48	38	45	33	38	42	67	40	34	39	42	46	52	44	46	42	41	40	35
Plumbing	14	47	33	40	43	31	38	46	67	44	51	34	40	46	46	43	49	44	44	40	33
Sheet metal	6	54	39	34	43	40	38	44	43	40	63	42	44	47	32	40	43	47	48	45	37
Woodworking	23	44	46	49	44	35	35	40	45	38	31	37	42	38	52	39	52	43	40	39	38
Electrical, general	105	47	27	51	44	34	38	43	34	34	33	40	40	36	50	46	44	43	41	39	51
Electrical, radio/TV	31	48	23	52	45	39	35	43	30	34	33	42	41	36	52	40	39	43	41	43	58
Electrical, industrial	19	47	25	51	44	41	45	44	30	36	37	37	39	40	46	46	38	42	43	39	50
Foods, baking	4	80	33	25	51	44	20	59	45	38	25	32	47	29	62	45	53	54	68	47	30
Foods, quantitative	4	80	33	35	49	34	18	54	36	44	23	46	49	24	64	42	50	49	70	45	35
Machinist	90	46	37	32	43	43	38	40	38	38	39	38	42	44	50	39	48	43	42	43	35
Welding	8	50	33	43	42	32	32	43	34	39	39	40	45	41	54	44	56	50	44	48	40
Mechanic, auto	64	52	34	39	45	36	39	44	40	39	34	38	40	39	48	46	48	45	46	41	39
Mechanic, aviation	40	46	33	43	45	44	43	40	34	38	39	38	42	40	50	40	40	43	45	44	44
Mechanic, marine	5	42	37	45	38	23	33	55	37	34	42	24	49	40	41	51	44	41	40	36	41
Printing	19	51	21	37	50	31	20	50	33	31	20	55	59	23	59	34	44	49	43	43	45

Note.—Standard scores are bold-faced to indicate the most relevant (criterion) scales for each of the student groups.

scores should be made from Clark’s stand-
ards, where a score of 40 on each scale con-
stitutes the mean for “Tradesmen-in-General”
(TIG). Scores of at least 50 or 60 suggest
similar or strongly similar interests to adults
employed in the occupation represented by
each scale. A score of 60 on one of these
scales means that a student’s score falls at
the mean of the criterion group. Scores of
30 and below indicate quite dissimilar inter-
ests with persons in these occupations. Those
MVII scales which are closely related, or
relevant, to each trade group have had their
corresponding mean scores bold-faced in the
table.

Several trends among these average inter-
ests scores are apparent: both high and low
scores help describe characteristic profiles for
some of the trades; some, but not all, of the
trades score highest on their own relevant
scales; almost all of the trades score moder-
ately high on some scales—e.g., baker. Inter-

pretative caution is advised since several of
the groups in Table 1 have small Ns.

The food, electrical, and printing trades
have the most valid profiles at the twelfth-
grade level. Both food groups, although very
small, show their highest score on the baker
scale and their second highest score for food
service manager. Scoring high on both food
interest scales is consistent with Scott’s find-
ing (1960) for his Minneapolis sample where
he reported the food service manager key
was better for uncovering food interests than
the baker key. The Buffalo food groups score
moderately high on clerical interest (IBM
operator and stock clerk) and personal serv-
ice scales (milk wagon driver, hospital at-
tendant); their lowest scores are on the
technical or manual trade interest scales.

The electrical groups are the only trades
to average at least moderately high scores
on electrician and radio-TV repairman, the
two criterion scales in this instance. Boys in

the radio-TV program score highest on this one; their average score of 58 is practically identical with the score of 60 for adults in the criterion sample. Scores on the building trade interest scales typically are low for these electrical students, even with the "general electrical" students, who are the most likely persons later to be involved in construction work.

The single printing trade group stands out for its many low scores. These students score moderately high (close to the criterion group score of 60) on the relevant pressman and printer scales; elevated scores also appear on such nontechnical, or manual, scales as: baker, IBM operator, milk wagon driver, and stock clerk.

Guidance uses of the MVII at the Grade-12 level, at least with this student sample, are not always encouraging. With the two food groups, we have a clear "hit." The two relevant scales clearly pick out these students and sharply focus on their trade interests. The students in the printing curriculum are also clearly spotted; on both the printer and pressman scales they are close to the criterion score of 60. The electrical students are spotted with only fair accuracy—a "near hit." All the three student groups are correctly indicated by the two MVII scales for electrician and radio-TV repairman. The best job of separation is done with the most specialized student group, that of radio-TV. For all these student groups, then, the MVII has clear guidance utility.

The situation is clearly otherwise with the students in various building trade programs as is also the case with student machinists, mechanics, welders. Here we have a clear "miss"; students in these programs are frequently found to earn their highest MVII scores on noncriterion scales. This needs to be checked out on additional student samples.

Similar results, although with fewer hits, are found for the ninth graders. Tabular data, as for the Grade-12 sample, will not be presented here. The majority of the scores hover around the TIG score of 40 which would suggest that the trade interests of these students are not as yet well differentiated. This is hardly an unexpected finding considering the lack of positive evidence for vocational ma-

turity found (in a different student sample, to be sure) in the Career Pattern Study (Super & Overstreet, 1960).

At this Grade-9 level, students in building trades and electrical groups are labeled as unclassified, a reflection of Buffalo vocational high school policy to expose these students at the Grade-9 level to a general orientation sort of course. It is not until their second year that students then divide up into specialties within these areas. If we disregard the generally high baker score for most of the groups—a finding also reported by Scott (1960)—the MVII profile for the unclassified building trade group is flat; that for the unclassified electrical group is better in the sense that the students here are showing up with modestly elevated scores on the two criterion scales: electrician and radio-TV repairman. The two food groups are sharply differentiated by means of both of the critical scales. Even should we choose to disregard the baker scale, these groups are spotted by the milk wagon driver scale and, less well, by the food service manager scores. The other trade groups (mechanics, machinists, tool designers) are poorly differentiated, the majority of all MVII scores clustering around the mean of 40.

It appears that many students at the ninth-grade level have vague interests or are undecided about their eventual trade. The tendency for so many boys to score higher on nonmanual or technical scales such as baker, stock clerk, and warehouseman, and lower on the carpenter, mechanic, plumber, and sheet metal worker scales, suggests a focus on "clean" work involving tangible operations. This trend appears at both the ninth- and twelfth-grade levels.

Clark (1961) presents data, based on the achievement records of naval personnel in aviation machinist training, which suggests that when learning ability is just adequate (i.e., at the average level), the motivational aspects of interests may then play an important role in school achievement. In his Table 35 (p. 89) he notes that the highest correlation between school grades and interest scores occurs with General Classification Test scores of 60–62; above and below this

narrow range, the size of the correlations drops radically.

We tried to check this with our twelfth grade vocational high school boys. Here we separated the curriculum groups into "high," "average," and "low" in terms of the aptitude test data. High was arbitrarily defined as IQs of 110 and above, average as 91-109, and low as 90 and below. For the DAT subtests (as Verbal, Spatial, Mechanical) high was set at 80 percentile or above for Form A (ninth-grade norms), low for the bottom 20%, average in the middle range of 21-79 percentile. Academic and shop course school averages were defined as: high of 85 and above, low as 69 and below, with the average range as 70-84. We then calculated the average MVII standard scores on only the criterion scales for these different ability groups as defined either by IQ or a particular DAT subtest or their school average. Data are presently being analyzed for only the large student groups. The first analysis done concerned the entire electrical group (general, radio/TV, industrial as shown in Table 1). None of the psychological test data (IQ or DAT subtests), when grouped into high, average, and low score ranges, produced differences for MVII criterion scale scores. When we look at these same students, this time classified by means of their academic and shop school averages (using these as our achievement measure), we then find the high students earn higher MVII criterion scale scores and that students with achievement indices below this level typically earn lower MVII criterion scores. We thus find a relationship between interest and achievement but not limited merely to average mental level. These data are presented as merely suggestive. The junior author is currently analyzing other large curricular groups.

SUGGESTED MVII REVISIONS

Clark (1961, p. 19) mentions the fact that some difficulties have arisen with the inventory because of vocabulary level. He further states that, if the inventory were to

be rewritten, a vocabulary level of a lower order would be utilized. From our experience with these vocational high school boys, both at the ninth- and twelfth-grade levels, we would heartily concur with this suggestion. There are far too many "advanced" technical terms employed in the inventory; much administration time was spent in explaining to the boys the meanings of such terms (examples, smear, chef, synthetic, sociology, personnel, comptometer). When publishing plans are finally agreed upon, it would therefore be wise to develop both an adult and a student form, the latter geared to a vocabulary level considerably reduced from the present one.

A reduced version of the MVII in another sense would also be helpful for this proposed student form. The inventory in its present form is too long for use with young boys, especially as early as Grade 9. Most of the boys needed more than 1 hour to complete the task; when even small amounts of time are added to this because of additional biographical data collection, one easily runs into a 1½- or 2-hour testing session. With smaller groups, and utilizing class time, two administration sessions might be planned and with more reliable results.

REFERENCES

- CAMPBELL, D. P., & SORENSON, W. W. Response set on interest inventory triads. *Educ. psychol. Measmt.*, 1963, 23, 145-152.
- CLARK, K. E. *Vocational interests of nonprofessional men*. Minneapolis: Univer. Minnesota Press, 1961. P. 129.
- MOTTO, J. J. Interest scores in predicting success in vocational school programs. *Personnel Guid. J.*, 1959, 37, 674-676.
- SCOTT, T. B. Counseling interpretations for the Minnesota Vocational Interest Inventory based on comparisons with the Strong Vocational Interest Blank. Unpublished doctoral dissertation, University of Minnesota, 1960.
- SUPER, D. E., & OVERSTREET, P. L. *The vocational maturity of ninth grade boys*. New York: Bureau of Publications, Teachers College, Columbia University, 1960. P. 212.
- WOLFEIN, S. L. Outlook for the skilled worker in the U. S.: Implications for guidance and counseling. *Personnel Guid. J.*, 1961, 39, 334-339.

(Received September 20, 1963)

THE EFFECTS OF TASK AND METHOD OF STIMULUS PRESENTATION ON THE DETECTION OF DECEPTION

LAWRENCE A. GUSTAFSON AND MARTIN T. ORNE¹

*Institute of the Pennsylvania Hospital and University of Pennsylvania*²

In a detection of deception experiment comparisons were made of the effects of 2 methods of stimulus presentation and 2 different subject tasks. The relevant-irrelevant method of stimulus presentation proved equally effective for both tasks, but the peak-of-tension method was significantly less effective where the S's task was to deceive as to the nature of guilty information possessed (guilty information paradigm) than it was where the task was to deceive as to the possession of any information (guilty person paradigm). In general, Ss found it easier to deceive in the guilty information paradigm, where they could attempt to "appear guilty" on a noncritical item and especially when they could anticipate the order of presentation of items (peak-of-tension method).

This study was designed to investigate the relative effectiveness of two different methods of stimulus presentation currently used in the detection of deception. The method most used in experimental studies presents significant items randomly interspersed with irrelevant ones without the subjects (Ss) knowing which items will be presented next. This technique, once widely used in commercial lie detection, is known as the relevant-irrelevant method (RI). The method of presentation which has been most popular in recent years in commercial lie detection was described in detail by Inbau and Reid (1953) and by Lee (1953) and is called the peak-of-tension method (PT). Here stimuli are presented to S in a known sequence. He knows exactly when crucial items will be presented and the technique is built upon this knowledge. The operator looks for gradual increase in tension coming to a peak at the crucial item with a sudden relaxation of tension when the item is past.

¹The research in this study was supported in part by the Institute for Experimental Psychiatry and by the United States Army Medical Research and Development Command Contract DA-49-193-MD-2480.

²The authors wish to express their appreciation to Emily Carota Orne and M. J. Moskowitz for their critical comments in the preparation of this manuscript. This research was conducted at Massachusetts Mental Health Center and Harvard Medical School.

In careful analysis of our pilot study data it became clear that there may be a marked interaction between these two methods of stimulus presentation and the precise task which is required of S in order to deceive. In most laboratory studies S is given an item of information of five or six possible items and the experimenter (E) studies the differences in physiological responses to these items to detect the "guilty information" (this will be called the "guilty information paradigm").³ Previous findings by the authors indicate that successful detection depends on S's motivation to deceive (Gustafson & Orne, 1963). Careful postexperimental inquiry of the motivated Ss and analysis of the data indicate that Ss who successfully deceive most frequently do so by volitionally producing a physiological response to the wrong item rather than by suppressing the "automatic" response to the crucial one. In the laboratory situation designed to study the guilty knowledge this strategy effectively prevents detection. However, it would not work in most field situations where the

³The similarity between the guilty information paradigm used here and the "guilty knowledge technique" used by Lykken (1960) should be pointed out. The authors acknowledge their debt to Lykken in the formulation of this paradigm. Lykken (1959) has used the guilty person paradigm, but in neither case does he point out the differences in the dynamics of the two situations.

suspect's task is to convince the operator that he is innocent—not that he is guilty of something else. This situation, which may be called the “guilty person paradigm,” requires *S* to mask all his responses rather than voluntarily responding to an irrelevant item.

For these reasons an experimental guilty person situation was designed. Whereas in the guilty information situation *S* selects one card from a deck where all the cards are number cards with the task of *E* to determine which was chosen, in the guilty person situation, *S* selects a card from a deck in which there are both blank and number cards with *S*'s task to convince *E* that he has drawn a blank card. In the guilty information situation it is thus possible for *S* to employ the strategy of creating false positives and succeed in deception while in the guilty person situation this strategy could not succeed. In the latter the *S* is required to appear innocent, i.e., to have drawn a blank card, and success is defined as: “If you draw a blank card, you must appear innocent—as in fact you are. If you draw a number, you must still appear as though you had drawn a blank card, i.e., innocent. If you seem to have drawn a numbered card whether or not you did, you fail on the experiment.” This now more closely approximates real life where the suspect must appear innocent to the operator in order to succeed. It does not help him to appear guilty of crimes he did not commit, nor does he serve his needs to appear guilty if he is in fact innocent.

The present study was an attempt to make a direct comparison of the detection of deception between the two paradigms of deception discussed above and using the two methods of stimulus presentation described. Specifically, the study was designed to determine the relative effectiveness in detection of the two methods of stimulus presentation (PT and RI) when used with two models of deception (guilty information paradigm and guilty person paradigm).

METHOD

Subjects. Fifty-three male undergraduate students from local universities were recruited from their school employment offices and paid for their participation. They had not previously taken part in

studies of detection in deception, but many had taken part in other psychological experiments.

Procedures. The *Ss* were divided into two groups with 24 *Ss* in the guilty information group and 29 *Ss* in the guilty person group. (Because in the guilty person group it was possible for a *S* to be “innocent” on one or more trials, extra *Ss* were run until a group of 24 had been obtained who were guilty on both a PT and an RI trial.)

Each *S* in both groups received two trials using PT (peak-of-tension) presentation and two trials using RI (relevant-irrelevant) presentation of materials.

The stimulus materials consisted of packs of seven cards each bearing a different number (for the guilty person group, two of the seven cards in each pack were blank) and corresponding tapes on which *E* had vocally recorded the same numbers, each number being presented twice. For different trials the numbers on the cards and tapes were of different series, the series used being 22–28, 32–38, 42–48, and 52–58. For PT trials, the numbers were recorded in ascending order and repeated in descending order. For RI trials, the numbers were presented in random order and repeated in a different order.

The Guilty Information Group. Each *S* listened to a tape recording of instructions which informed him that his task was to try to deceive *E* as to the number on the card he picked at the beginning of each trial. The *S* was instructed to write the number he had picked on a sheet of paper which would later be sealed in an envelope to insure that he actually attended to the number. He was also told that numbers would be read to him including the number which he had selected, that these numbers would sometimes be read to him in numerical sequence and sometimes in random order, and that he would be informed at the beginning as to the method of presentation of the numbers. Both *E* and *S* knew that all cards had numbers on them.

When the instructions had been given, *S* was given an opportunity to ask questions about the procedure which *E* answered when possible. The *E* then attached the necessary electrodes for recording skin resistance, heart rate, respiration, and finger pulse volume. The methods for recording skin resistance were those used by Wenger, Engel, and Clemens (1957). (The other measures were included for exploratory reasons and will not be discussed further.)

All *Ss* then drew a card, looked at it, wrote the number down on another card, and sealed it in an envelope. The *E* then went into an adjoining room and turned on the polygraph and the tape recording, which told *S* whether the numbers would be in consecutive order or in random order. The numbers were presented at 15-second intervals. At the completion of one trial another card was drawn from a different range of numbers and the procedure was repeated.

The Guilty Person Group. Each *S* in this group was given taped instructions similar to those given

to the first group. Here, however, the *S* was informed that at times (probably on at least one trial) he would draw a blank rather than a number card and that his task on all trials was to try to convince *E* that he had drawn a blank rather than a number. When a blank was drawn (*S* was instructed), he was to write a zero on the slip of paper. Except for these instructions, each *S* was treated in the same manner as the first group.

The *Ss* in both groups were instructed that they were to say "no," as they heard each number, in such a way that their voice would not give them away. No information was given between trials to *Ss* concerning how successful they had been at deception. Each *S* was interviewed after the session to see how the *S* perceived the situation. It was especially important that the *Ss* in the guilty person group perceived their task as that of trying to make *E* think they had drawn a blank card and not another number.⁴

ANALYSIS OF THE DATA

The basic response data analyzed in this study were the skin resistance and changes in skin resistance as recorded by the GSR. As in the previous study by the authors, the primary concern in this analysis was not with the number of correct detections per se, but rather with the effect of the instructions and experimental procedures on the response to critical items relative to the response to other items.

Because of the differences in the dynamics of the RI and PT methods of presentation, different objective criteria of deception were required. In the RI trials, the magnitude of change in skin resistance was measured for each stimulus (number). The stimuli were then ranked in terms of average response magnitude, the one showing the greatest average change being ranked one, that showing the second greatest change being ranked two, etc. Once all stimuli in a trial had been ranked, the rank assigned to the critical item for that trial could be determined and used as the measure of relative responsiveness for purposes of analysis.

On PT trials, *S* anticipates the stimuli prior to their presentation, so that the prestimulus response level is affected by the stimulus. Thus magnitude of change in resistance from the prestimulus level is not an effective measure of responsivity. For this reason, the

lowest level of skin resistance for each stimulus in a trial was taken as the criterion. The mean level for each stimulus (number) was calculated and a rank of one assigned to the stimulus producing the lowest level, etc. As in the other method, the rank assigned to the critical item in the trial was determined and used in the analysis.

Since comparisons of the relative effectiveness of RI and PT methods of stimulus presentation are intrasubject comparisons, the Wilcoxon matched-pairs signed-rank test was used here. For comparisons of the relative effectiveness of the guilty information and guilty person groups the Mann-Whitney *U* test was used.

As it was possible for *S* to draw a blank card in the guilty person paradigm on one or more trials, some *Ss* were not guilty on both a PT and RI trial. Therefore, it was necessary to discard such *Ss* in order to make intrasubject comparisons. Of the 29 *Ss* run, 24 were guilty on both RI and PT trials. Only these 24 are included in the analysis presented here.

RESULTS

As can be seen in Table 1, detection in the guilty person paradigm proved significantly superior to that in the guilty information paradigm (overall Mann-Whitney *U* = 147, $n_1 = n_2 = 24$).

TABLE 1
COMPARISON OF RI AND PT METHODS OF STIMULUS PRESENTATION IN GUILTY PERSON PARADIGM AND GUILTY INFORMATION PARADIGM EXPERIMENT.

	Mean ranks		<i>p</i>
	Guilty information	Guilty person	
PT	2.41	1.33	<.01 ^a
RI	1.69	1.27	>.50 ^a
<i>p</i>	<.05 ^b	>.05 ^b	
Number of times the critical number was assigned a rank of one.			
PT	8	19	<.01 ^c
RI	15	18	>.50 ^c

Note.—*N* = 24 in each group.

^a Mann-Whitney *U* test.

^b Wilcoxon matched-pairs signed-rank test.

^c χ^2 test.

⁴ All subjects in the guilty person group did try to do this.

There was no significant difference in the effectiveness of the RI method of stimulus presentation between the guilty person paradigm and the guilty information paradigm. On the other hand, the PT method proved significantly less effective than the RI method in the guilty information paradigm, and significantly less effective in that paradigm than it was in the guilty person paradigm, both in the distribution of ranks assigned to the critical items and in the number of correct detections (the number of times a rank of one was assigned to a critical item). (There was no difference in the effectiveness of the PT and RI methods in the guilty person paradigm, both proving very effective.)

The use of nonparametric tests of significance, required by the nature of the data, makes it impossible to test directly for interaction effects. However, the data of Table 1 provide rather convincing evidence for an interaction between paradigms and methods of stimulus presentation as described in the preceding paragraph.

DISCUSSION

The results of this study and those of the preceding study of this series suggest several points in two major aspects of the detection of deception, the first dealing with the results of experimental studies in this area and the second dealing with the behavior of the deceiver.

In the present study, the RI method of stimulus presentation proved more effective in detection than the PT method when *S* was trying to deceive *E* as to which item of information he possessed (guilty information paradigm). The RI method and the guilty information paradigm is the combination most commonly used in laboratory studies of deception (Burt, 1921; Ellson, Davis, Saltzman & Burke, 1952; Kubis, 1962; Lykken, 1960; Marston, 1917; Van Buskirk & Marcuse, 1954; as well as Gustafson & Orne, 1963). Thus it would appear that these studies have used the best method of stimulus presentation for the type of deception that they were studying.

However, laboratory investigations differ in a variety of ways from the field situation. One is "real life" and the other is an "ex-

periment." Some have felt that this is an overwhelming obstacle. However, it is possible to investigate variables which will be working in both situations. The laws of behavior are not different in the two situations. Unfortunately, many of the laboratory studies do not take into account variables which are operating in a field situation and as a result use designs which are not generalizable to the field situation. This is likely to lead to false generalizations about the field situation, when in fact the problem is in the experimental design itself.

For example, the relative ineffectiveness of the PT method, as reported in this study, for the guilty information paradigm (the typical laboratory situation) would seem to make it a poor choice in the detection of deception. However, when the conditions more closely approximated the field situation, i.e., when the explicit task of *S* is to appear as though he did not have the information (guilty-person paradigm), the PT method was no less effective than the RI method.

Two recent studies have used experimental situations similar to the guilty person paradigm of the present study. One of these, Kubis (1962), reports successful detections of the order of 97% in one condition which used judgments of experienced operators. Lykken (1960) using only objective methods of classification reported 90% success in detection. Compared to the 70% success reported in laboratory studies using the guilty information paradigm (see, for example, Ellson et al., 1952), the results of these studies support the findings of the present study that the guilty person paradigm provides significantly greater ease of detection.

In general, then, it is necessary to urge caution in the comparison of laboratory studies of deception where different paradigms and methods may be used and still greater caution in the application of findings obtained in the laboratory to situations occurring in the field.

In order for deception to be detected, the responses of the deceiver to critical items must differ from his responses to noncritical items. The present study and its predecessor indicate two factors which affect this differentiation: the motivation of the deceiver and the tactics which he must use to deceive.

The differentiation of responses to critical and noncritical items can be heightened in two ways: increasing the magnitude of response to the critical items or reducing the amount of response to noncritical items (background activity). Increasing the motivation to deceive increased the response to critical items over that to noncritical items (Gustafson & Orne, 1963). Background activity was also increased; in other words, Ss showed increased responsiveness in general. This increase in general activity is probably due to the use in that study of the RI method of stimulus presentation which requires S to maintain a "set" for the critical item which can occur at any time.

In the PT method of presentation, where S can correctly anticipate the occurrence of the critical item, it would be expected that background activity would be reduced. The present study supports this expectation. There were significantly fewer responses overall during PT trials than during RI trials (Wilcoxon matched-pairs signed-ranks test: $T = 112.5$, $n = 28$, $p < .05$).

As is suggested by the present study, the tactics of the deceiver vary with the paradigm and simultaneously with the means of stimulus presentation. (The interaction between paradigms and methods of presentation has been pointed out in the results of this study.) In the guilty information paradigm the deceiver can, and usually does, attempt to shift the attention of E to some noncritical item—an attempt which is easiest when S is able to anticipate the occurrence of the items (PT method). It is this situation, as was found in this study, which makes deception most difficult to detect.

The results of the comparisons of the relative effectiveness of different paradigms and different methods suggest that it is easier for S to produce autonomic nervous system (ANS) responses (at least of small magnitude) to noncritical items than it is for S to inhibit responses to critical items, an observation commonly made at the clinical level. This is an important point for discussions of "control" of autonomic responses.

It would appear, then, that the "optimum" conditions for detection of deception would

be found in a situation where S must prove that he is innocent (the guilty person model) where he is very highly motivated to deceive (heightened response to critical items), and where he "knows" exactly when he must deceive (decreased background activity with PT presentation).

Finally, the results of these studies indicate that the detection of deception is not a simple matter of asking a laboratory deceiver (or suspect) a few questions and recording his physiological response. It is an extremely complicated process which appears to be a function of a variety of psychological factors. Fortunately, however, these factors, as well as the physiological procedures themselves, are subject to experimental investigation.

REFERENCES

- BURTT, H. E. The inspiration-expiration ratio during truth and falsehood. *J. exp. Psychol.*, 1921, 4, 1-23.
- ELLSON, D. G., DAVIS, R. C., SALTZMAN, I. J., & BURKE, C. J. A report on research on detection of deception. Contract Nonr 60nr-18011, 1952, Indiana University, Department of Psychology, Office of Naval Research.
- GUSTAFSON, L. A., & ORNE, M. T. The effects of heightened motivation on the detection of deception. *J. appl. Psychol.*, 1963, 47, 408-411.
- INBAU, F. E., & REID, J. E. *Lie detection and criminal investigation*. (3rd ed.) Baltimore: Williams & Wilkins, 1953.
- KUBIS, J. F. Studies in detection: Computer feasibility considerations. Technical Report No. 62-205; June 1962, Rome Air Development Center, New York; Contract AF 30(602)-2270, Project 5534; United States Air Force, Armed Services Technical Information Agency (now known as Defense Documentation Center), No. AD-284 902.
- LEE, C. D. *The instrumental detection of deception*. Springfield, Ill.: Charles C Thomas, 1953.
- LYKKEN, D. T. The GSR in the detection of guilt. *J. appl. Psychol.*, 1959, 43, 385-388.
- LYKKEN, D. T. The validity of the guilt knowledge technique: The effects of faking. *J. appl. Psychol.*, 1960, 44, 258-262.
- MARSTON, W. M. Systolic blood pressure symptoms of deception. *J. exp. Psychol.*, 1917, 2, 117-163.
- VAN BUSKIRK, D., & MARCUSE, F. L. The nature of errors in experimental lie detection. *J. exp. Psychol.*, 1954, 47, 187-190.
- WENGER, M. A., ENGEL, B. T., & CLEMENS, T. L. Studies of autonomic response patterns: Rationale and methods. *Behav. Sci.*, 1957, 2, 216-221.

(Received September 20, 1963)

JOB CHARACTERISTICS AS SATISFIERS AND DISSATISFIERS

FRANK FRIEDLANDER¹

United States Naval Ordnance Test Station, China Lake, California

It is often assumed that job satisfaction and dissatisfaction are opposites, and that one is the mere negation of the other. This assumption of convertible bipolarity is examined by administration of 2 questionnaires to 80 Ss in which the importance to satisfaction and the importance to dissatisfaction of various job characteristics are compared. Correlational and variance analyses both indicate that satisfaction and dissatisfaction are, for the most part, unrelated and not complementary functions, rather than negatively related poles of a single bipolar continuum. Results of studies and theories utilizing a single satisfaction-dissatisfaction continuum are thus questionable. Summary data of ranks of satisfiers and dissatisfiers are discussed in regard to current job motivation theory.

Inherent in most devices which purport to measure job attitudes is the assumption that job satisfaction and job dissatisfaction are opposites, i.e., that if we measure the extent of an employee's satisfaction with various aspects of his job and his organization, we have concurrently also measured his lack of dissatisfaction. We usually assume that job attitudes are on a bipolar continuum and that the more satisfied an employee, the less dissatisfied he is. Bipolarity, in this sense, refers to the assumption that the attitude extends from one extreme of satisfaction, through zero, to an extreme dissatisfaction, and that the two kinds of behavior are negatively correlated. To the extent that this assumption is valid, there is some justification to convert the scale (and the construct) to a unidirectional one. The danger of converting attitude scales which are inconvertible into unidirectional continua is discussed by Thompson (1961); such dangers, however, appear to be ignored in most research on job attitudes.

Since assumptions of convertible bipolarity underlie the tools for measuring attitudes, these assumptions also underlie the theories on job attitudes which have evolved from studies utilizing these tools. A survey of the literature indicates that nearly all measuring

instruments and resultant attitude theories are based on this premise. Recently, a study by Herzberg, Mausner, and Snyderman (1959) tended to cast some doubt on the conventional assumption of a bipolar satisfaction-dissatisfaction continuum. However, Schwarz (1959) found indications that dissatisfaction results primarily from the prevention or impairment of achieving satisfaction. As might be expected from a new mode of analysis, results of the few studies performed are somewhat ambiguous.

The purpose of this study was to subject the assumption of a bipolar continuum to quantitative analysis. Specifically, several questions were asked. Will respondents who find certain characteristics of the job particularly satisfying also experience pronounced dissatisfaction when these elements are lacking in the job? According to conventional assumptions, we would expect a job characteristic which serves as a strong satisfier to also serve as a strong dissatisfier if the characteristic is lacking or negative in the job. Peripheral to this core query is the question of which job characteristics provide the greatest source of satisfaction, and which provide the greatest source of dissatisfaction?

PROCEDURE

In accordance with the questions posed, two separate questionnaires were used. One measured the importance ascribed by the respondent to each of 18 variables as sources of satisfaction, and the second

¹ This study was performed, in part, while the author was on the staff of Psychological Research Services of Western Reserve University.

measured the importance ascribed by the same respondent to the lack of or negative aspect of the same 18 variables as sources of dissatisfaction. Directions for the first of these measures were as follows:² "Think of a time when you felt exceptionally good about your job, either your present or any other job you have had. The following is a list of some of the factors which may have contributed to your good feeling at that time. How important was each of these factors in the particular experience you are describing?" Directions for the second measure were identical except the word "dissatisfied" was substituted for the word "good." The items for both the satisfaction and dissatisfaction measures have been deposited with ADI.³

The satisfaction measure was identical (with one item added) to that used by Friedlander (1963) in a factor analytic study. In both the satisfaction and dissatisfaction measures, the respondent checked one of four blanks extending from lack of the job characteristic as a contributor to his satisfaction (dissatisfaction) to its major importance in the experience he was describing. The respondent thus was not questioned as to whether he was satisfied or dissatisfied; rather he was asked to draw upon his entire past vocational repertoire and to indicate the extent to which each job aspect was important as a source of satisfaction and dissatisfaction.

Over half of the 80 subjects (Ss) were full-time employees in a variety of occupations and positions and were attending an evening course in either industrial psychology or child psychology. The remaining respondents were attending day classes at a college which utilizes a cooperative work program; these latter Ss thus worked and then attended college alternately for 3-month periods. The Ss ranged in age from 20 to 50 and had a mean age of 25. The range in full-time equivalents of work experience was from 1 to 34 years and the mean was 7 years.

The Ss were administered the two measures 1 week apart at the beginning of whichever of the above courses they attended. There is no reason to suspect that Ss were not naive as to the specific purposes of the questionnaires. The order in which the two measures were administered was reversed in half of the administrations, and since no significant differences existed between the results of administrations, there is no cause for considering the sequence of administration of the two measures had any appreciable effect upon the response.

² The satisfaction measure was designed, in part, by Frederick Herzberg of Western Reserve University.

³ The two questionnaires, the two analyses of variance tables, and the two Multiple Range Test tables have been deposited with the American Documentation Institute. Order Document No. 8027 from ADI Auxiliary Publications Project, Photoduplication Service, Library of Congress, Washington, D. C. 20540. Remit in advance \$1.25 for microfilm or \$1.25 for photocopies and make checks payable to: Chief, Photoduplication Service, Library of Congress.

RESULTS

The reliabilities of the satisfaction and dissatisfaction measures were computed by means of the Kuder-Richardson Formula 20, and were .79 and .72, respectively. The remainder of the results will be discussed under two subheadings in accordance with the two general questions outlined earlier: (a) a comparison between satisfiers and dissatisfiers; and (b) a comparison among the several satisfiers and, separately, a comparison among the dissatisfiers.

Comparison between Satisfiers and Dissatisfiers. For the purpose of comparing satisfiers and dissatisfiers, the data were subjected to two different types of analyses: (a) a test of the significance of the difference between a job characteristic as a source of satisfaction and the lack or negative aspect of the job characteristic as a source of dissatisfaction, and (b) calculation of the degree of relationship between the job item as a satisfier and as a dissatisfier by use of Pearson product-moment correlations. Table 1 is a summary of the data obtained. In Columns 1 and 2 are listed the means and standard deviations of the 18 job characteristics as satisfiers; similar data is listed in Columns 3 and 4 for dissatisfiers. The difference between these two means is indicated in Column 5; an asterisk in Column 5 indicates a significant difference beyond the .01 level between Columns 1 and 3.

If satisfaction and dissatisfaction are complementary functions, one would expect no significant differences between the mean satisfaction and mean dissatisfaction score for the same item. For 12 of the 18 job characteristics, satisfaction with the job item differs significantly from dissatisfaction with the lack of or negative aspect of the item. Thus, for most job characteristics, satisfaction and dissatisfaction are not complementary functions. Good working conditions, for example, are not equally important to satisfaction as poor working conditions are to dissatisfaction. In fact, for all job items in which the difference is significant, the item serves far more as a source of satisfaction than as a source of dissatisfaction. This tendency is further substantiated by the fact that the mean of all

TABLE 1
MEANS, STANDARD DEVIATIONS, AND CORRELATIONS BETWEEN SOURCES OF JOB
SATISFACTION AND SOURCES OF JOB DISSATISFACTION

Job item	1	2	3	4	5	6
	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>		
	Satisfaction		Dissatisfaction		Difference	<i>r</i>
Promotion	2.10	1.17	1.99	1.13	.11	.34*
Challenging assignments	3.10	.99	2.67	1.10	.43*	-.05
Recognition	3.23	1.01	2.61	1.11	.62*	.08
Relations with supervisor	3.06	.97	2.09	1.17	.97*	-.14
Relations with coworkers	2.73	.94	1.46	.89	1.27*	-.19
Technical supervision	2.56	1.11	1.75	1.06	.81*	-.03
Merit increases	1.80	1.09	1.60	1.01	.20	-.10
Achievement	3.38	.86	2.80	1.10	.58*	-.12
Working conditions	2.31	1.08	1.60	.94	.71*	-.12
Responsibility	2.84	1.10	2.01	1.07	.83*	-.16
Security	2.56	1.00	1.64	.99	.92*	.14
Growth	3.38	.89	2.19	1.19	1.19*	-.11
Employee benefits	1.46	.91	1.30	.70	.16	.00
Work itself	3.46	.72	2.39	1.31	1.07*	-.12
Home life	1.63	1.08	1.61	.97	.02	.30*
Work group	2.19	1.09	1.77	1.00	.42*	-.02
Management policies	2.01	1.02	2.28	1.17	-.27	-.16
Use of best abilities	2.99	1.02	2.89	1.14	.10	-.31*
<i>M</i>	2.60		2.04		.56*	
<i>SD</i>	.61		.47			

Note.—*N* = 80.
* *p* < .01.

satisfaction items is significantly greater than the mean for all dissatisfaction items.

Column 6 in Table 1 indicates the degree of relationship between the importance of the job characteristic to satisfaction and to dissatisfaction. Fifteen of the 18 correlations are not significant, indicating that few accurate predictions of worker-item dissatisfaction can be made from a knowledge of the employee's specific satisfactions. Generally, to the extent that these items are important to satisfaction, lack of these may or may not be important to dissatisfaction.

In only three cases is the correlation significant in Column 6. The two positive correlations tend to indicate that to the extent that promotion and repercussions of job on home life are important to satisfaction, the lack of or negative aspect of these are important to dissatisfaction. For these two items only, are the assumptions of some bipolarity of the satisfaction-dissatisfaction continuum partially substantiated. For the remaining 15 job characteristics, nonsignificant correlations

indicate that satisfaction and dissatisfaction are unrelated, and thus not bipolar.

Comparisons among Satisfiers and among Dissatisfiers. Whereas the previous analysis compared Columns 1 and 3 in Table 1 horizontally, the second hypothesis is concerned with a vertical analysis of the means in Column 1 and a separate vertical analysis of the means in Column 3. So as to identify those job characteristics which serve as the greatest satisfiers, and similarly, those which serve as the greatest dissatisfiers, two analyses of variance were performed. *F* ratios in both cases clearly indicate significant differences among the various job characteristics as sources of satisfaction, and significant differences among the same job characteristics as sources of dissatisfaction (see Footnote 3).

In order to obtain a better understanding of which source-of-satisfaction means differ from each other, and similarly which source-of-dissatisfaction means differ from each other, Duncan's Multiple Range Test (Duncan, 1955, 1957) was applied separately to

each set of means. With a significance level of .01, and k (means) = 18, the minimum probability of finding no erroneous significant differences between the 18 means (the protection level) is .84 (Edwards, 1960).

The results of these tests indicate a variety of differences among job characteristics as satisfiers, and separately as dissatisfiers. Job characteristics such as achievement, challenging assignments, recognition, and the work itself were viewed as most important to both satisfaction and dissatisfaction. These seem to be involved in the work process itself, and to hold importance for the more intrinsic satisfactions and dissatisfactions to be derived from the job. Work characteristics least important to both satisfaction and dissatisfaction were employee benefits, merit increases, working conditions, effect of job on home life, job security, and the technical competence of the supervisor. These encompass the social and technical environment of the worker. Although the ranking by the entire group of employees of satisfiers and dissatisfiers is quite similar, as noted earlier in the case of only a relatively few job items is there any relationship between a job characteristic as a satisfier and as a dissatisfier.

SUMMARY AND CONCLUSIONS

This study has indicated that there are significant differences between the importance that an employee ascribes to various job characteristics as a source of satisfaction as opposed to these same job characteristics as a source of dissatisfaction. Respondents who find certain aspects of the job particularly important to their satisfaction may not find the lack of or negative aspect of this same characteristic particularly important to their dissatisfaction. Nonsignificant correlations between satisfiers and dissatisfiers tend to indicate that satisfaction and dissatisfaction are not on a bipolar continuum. Furthermore, the majority of characteristics would seem to be significant contributors to both satisfaction and dissatisfaction, although the way in which each characteristic serves such dual functions is often different and unrelated for satisfaction as compared to dissatisfaction.

Since there are strong indications that

satisfaction and dissatisfaction are not negatively related poles of a single continuum, it is probable that one is not justified in converting (by combining) the two constructs into a single scale or a single construct. A better understanding of the complexities of job motivation might be gained by utilizing two single unidimensional constructs or scales in gathering data on job attitudes.

When the mean job-item responses were correlated for satisfiers and dissatisfiers, a high positive rank order correlation (.71) resulted. Thus, summary data covering many job characteristics may indicate similar rankings for satisfiers and dissatisfiers; however, caution is advised in the interpretation of such data. This study has indicated that for each individual job characteristic, few such relationships exist.

The conceptual framework suggested by Herzberg et al., is relevant to this study in that it is concerned with the double dichotomy of satisfiers versus dissatisfiers, and intrinsic versus extrinsic job characteristics. Herzberg's findings that satisfiers and dissatisfiers were not opposite ends of a common set of dimensions were substantiated by the current study. He further found that satisfiers dealt mostly with indices of personal growth and self-actualization, while dissatisfiers involved the environmental and physical characteristics of the job. From such findings, one might infer that intrinsic job characteristics would be most important to satisfaction and quite unimportant to dissatisfaction, while extrinsic job characteristics would be most important to job dissatisfaction and unimportant to satisfaction. Only half of this framework was substantiated by the current study; intrinsic job characteristics were found to be important to both satisfaction and dissatisfaction, while extrinsic aspects were relatively unimportant as satisfiers or dissatisfiers. These results coincide more closely with those of Schwarz (1959), who found that critical incidents leading to negative job attitudes usually involved frustrating managers' attempts at self-actualization.

The contrasting effects of intrinsic and extrinsic job characteristics are given further operational significance in a study by Fried-

lander and Walton (in press), which indicated that characteristics of the work content and process serve primarily to elicit positive motivations in attracting the employee to remain with his organization, while characteristics of the work context and community serve primarily to evoke negative motivations in causing the employee to be sufficiently dissatisfied to leave.

Generalizations of these findings to groups of different and specific characteristics should be done with caution. The individuals in this study were enrolled in at least one college course; as such, they had shown some evidence of wanting to better themselves. Furthermore, workers who are older, have longer work experience, or belong to other status and occupational levels might well indicate different job characteristics as important to satisfaction and dissatisfaction.

REFERENCES

- DUNCAN, D. B. Multiple range and multiple F tests. *Biometrics*, 1955, **11**, 1-42.
- DUNCAN, D. B. Multiple range tests for correlated and heteroscedastic means. *Biometrics*, 1957, **13**, 164-176.
- EDWARDS, A. L. *Experimental design in psychological research*. New York: Holt, 1960.
- FRIEDLANDER, F. Underlying sources of job satisfaction. *J. appl. Psychol.*, 1963, **47**, 246-250.
- FRIEDLANDER, F., & WALTON, E. Positive and negative motivations toward work. *Admin. Sci. Quart.*, in press.
- HERZBERG, F., MAUSNER, B., & SNYDERMAN, BARBARA B. *The motivation to work*. New York: Wiley, 1959.
- SCHWARZ, P. A. *Attitudes of middle-management personnel*. Pittsburgh: American Institute for Research, 1959.
- THOMPSON, J. W. The bi-polar and unidirectional measurement of intelligence. *Brit. J. Psychol.*, 1961, **52**, 17-23.

(Received September 23, 1963)

TEACHER ACCURACY IN ASSESSING COGNITIVE VISUAL FEEDBACK FROM STUDENTS¹

JON JECKER,² NATHAN MACCOBY, HENRY S. BREITROSE,
AND ERNEST D. ROSE

Institute for Communication Research, Stanford University

The relative value of verbal and nonverbal cues in teacher accuracy in making judgments of student comprehension was tested. 67 new interns (not yet teachers), 59 inexperienced teachers, and 46 experienced teachers were each shown 20 short sound-film recordings of 10 students being taught. They rated student comprehension. 57 Ss heard both picture and sound, 60 heard sound only, and 55 saw only the picture. When sound is absent, all groups of teachers are inaccurate in judging student comprehension (about $\frac{1}{3}$ correct with $\frac{1}{4}$ the chance base). When sound is present, whether or not the picture is seen, all groups exceed chance (about $\frac{2}{3}$ correct), but do not differ significantly from each other. The need for research on teacher training in the observation and interpretation of nonverbal feedback from students is indicated.

Communicators must continually make judgments during the actual process of communication on how successfully they are communicating. In face-to-face communication settings, the communicator has information feeding back to him from the audience while he is communicating. In a two-person setting, immediate and recurring verbal feedback is often possible. When a small group is the audience, at least occasional verbal feedback may take place, although even here it is likely to be limited largely to particular audience members. In general, however, as the audience becomes larger, the communicator must base judgments of communication success on less and less immediate verbal feedback.

Similarly, teachers must continually make judgments, while teaching, on how successfully they are communicating. Student response is an important source of immediate information useful to the teacher in making such judgments. However, while a teacher is addressing a sizable class of students, verbal feedback is likely to be severely limited, and the teacher is largely dependent upon non-

verbal information from students, such as facial expressions and other bodily movements, when judging the effects of his communication.

Teachers often feel that they can accurately interpret facial expressions, and other nonverbal student behavior, and thus correctly assess the general level of the student's understanding of the subject matter being presented. To our knowledge, however, this feeling has not previously been put to a controlled experimental test. The purpose of the experiment reported here was to assess the relative value of verbal and nonverbal cues as they contribute to a teacher's accuracy in judging student comprehension.

Continuous feedback that can be matched against what a communicator has been attempting to get across tends to improve the effectiveness of the communication. However, it was our observation that such matching of feedback occurs far less often in nonverbal than in verbal exchanges. We therefore hypothesized that when teacher judgments of student comprehension are based on visual (nonverbal) cues from students, misperceptions of student comprehension are likely to occur more frequently than when judgments are based on verbal cues.

Our second interest in this study was to investigate the effects of both training and experience in teaching on the accuracy of judgment of student comprehension. Al-

¹ The research reported in this paper is part of a National Defense Education Act Title VII project entitled "Sound Film Recordings in Improving Classroom Communications."

² This investigation was carried out during the tenure of a Predoctoral Fellowship to the first author now at Kent State University from the National Institute of Mental Health, United States Public Health Service.

though we did not expect to obtain a high positive relationship between experience and success in judging student comprehension, we did expect to find some correlation.

METHOD

An investigation of nonverbal feedback to teachers was conducted in the following manner. Sound-film recordings, which provide a complete and direct record of both verbal and nonverbal behavior, were made of individual students while being taught. These films were then used as stimulus material in an experiment to investigate the extent to which people, either teachers or those who are training to be teachers, can correctly judge how well a student is comprehending the material being presented.

Preliminary Filming and Pretesting

The first procedural step was to collect the sound-film recordings of student-teacher interactions. This was done by making sound-film recordings of 10 junior high school (eighth or ninth grade) students being instructed individually in elementary algebra by one of the investigators in our studio. Four of these students were female, six male.

Our next task was to develop, via pretest, a usable scale for judging student comprehension. To implement this pretest, the experimenters (*Es*) designed what appeared to be a reasonable scale (Table 1). Five segments, or "clips," from the film of each of four students were chosen as stimulus material. The criterion for choosing each clip was that it be a recognizable question-answer sequence and that it fit one of the categories of the previously designed scale of student comprehension. These 20 clips were then put together into a continuous filmstrip.

Twelve pretest subjects (*Ss*) were shown this film, one clip at a time. Immediately after each clip, *S* was asked to use our rating scale to judge the amount of student comprehension in that clip. The *Ss* were encouraged to discuss the basis of their ratings. After rating all 20 clips, *Ss* were interviewed by *E* as to what modifications might be made in the scale to improve its utility. These pretest

TABLE 1
ORIGINAL SCALE FOR JUDGING STUDENT COMPREHENSION

Understands clearly	+3
Follows	+2
Follows with some confusion	+1
Nothing happening	0
Slightly confused	-1
Clearly puzzled	-2
Totally lost	-3

Note.—Scale used in pretest.

TABLE 2
REPRODUCTION OF REVISED STUDENT COMPREHENSION RATING SCALE AND SUPPLEMENTARY SCALES USED IN THE MAIN EXPERIMENT

Clip No. —	Booklet No.	—		
A. In this clip, how much is the student understanding? (Circle one number OR check the statement below the scale.)				
1 Understands nothing	2 Understands a little	3 Understands quite a bit	4 Understands completely	
—————Can't tell.				
*	*	*	*	
B. How confident are you that your judgment of how much the student is understanding is correct? (Circle one number.)				
1 Not at all confident	2	3	4	5 Totally confident
*	*	*		*
C. In this clip, how much GENUINE effort is the student making to understand? (Check <i>only</i> one.)				
—————A great deal of effort.				
—————Some effort.				
—————Very little effort.				
—————No effort at all.				
—————Can't tell.				

sessions were tape recorded and later content analysed to determine which revisions of the original scale to carry out. The revised scale (Table 2, Question A) was a four point one where one was labeled "understands nothing" and four was labeled "understands completely." An offset category was included allowing the rater to indicate that he could not make a judgment.

The Experimental Film

An experimental film was then constructed. Sixteen film clips were chosen on the basis of being recognizable question-answer sequences which fit categories of the revised rating scale. Each category of the rating scale was represented by four clips. Four additional clips contained no instruction and therefore fit the offset (can't tell) category.

Two clips were taken from each of the 10 instruction films. These clips were chosen as being the best example of the comprehension category assigned, on a random basis, to be taken from each film. The researchers used all possible contextual information to rate and select the clips. Only clips on which 100% researcher agreement was achieved were used in the final experimental film. Since the

rating categories were broad, the researchers felt quite confident of their own accuracy in selecting the clips.

These 20 clips were then divided into two sets of 10, with each instruction film (and thus each student) supplying one clip in each set. The order of presenting sexes was alternating, except for some adjacent male clips due to the six male-four female imbalance in the instruction films. Within the alternating sexes restriction, the order in which individual students were shown in each half of the film was randomly determined, thus randomizing the actual degree of comprehension shown at any point during each set.

Subjects

One hundred seventy-two Ss representing three teaching-experience levels were employed.³ Sixty-seven students who had just enrolled in the teacher training program at Stanford made up the least experience, or New Intern group. Fifty-nine students who were nearing completion of the same teacher training program but who had not yet taken their first paying job as teachers made up the moderate experience, or Old Intern group. Forty-six teachers, enrolled in summer courses at Stanford and having at least 2 years experience made up the highest experience, or Experienced Teacher group.

Experimental Treatments

Each experience-level group was subdivided into three experimental conditions (Table 3). Approximately one third of each experience-level group was shown the whole experimental film, i.e., both the visual image and the sound track. The sound track consisted largely of teacher instruction. This Picture plus Sound condition will hereafter be referred to by P + S. Approximately one third of each experience-level group heard only the sound track (Sound Only condition designated by S-O), and the remaining Ss in each experience-level group saw only the visual image of the experimental film (Picture Only condition designated by P-O).

The schedule called for running the experiment in six groups. Proportionate numbers of Ss from each experience-level group were preassigned to each experimental hour, and all Ss in 1 experimental hour were in the same condition, P + S, S-O, or P-O. Table 3 presents the experimental design with actual Ns for the various cells.

Procedure

When Ss arrived at the experiment, they were handed a printed booklet. The first page of this booklet was devoted to a photograph of the setting in which the instruction films were taken. The next

³ The authors would like to express their gratitude to Dwight Allen of the Secondary Education Project of the School of Education at Stanford for co-operation and helpful suggestions in recruiting subjects.

TABLE 3
EXPERIMENTAL DESIGN

Group	Picture plus sound	Sound only	Picture only	Marginal
New interns	N = 20	27	20	67
Old interns	N = 20	18	21	59
Experienced teachers	N = 17	15	14	46
Marginal	N = 57	60	55	Total N 172

20 pages were identical; each contained three questions with their respective scales to be answered in response to each of the 20 clips of the experimental film (see Table 2). The first of these (Question A) asked S for his estimate of the degree of student comprehension. The second (Question B) asked S to indicate on a five-point scale how confident he was of his answer to Question A. The third (Question C) asked S to estimate how much genuine effort the student in the clip was putting forth to understand what was being explained to him.

Instructions to the groups were brief. A short explanation of the investigators' interests was given. The photograph on the first page of the booklet was used to familiarize Ss with the situation in which the films were taken. The various rating scales in the booklet were discussed in detail and instructions were given concerning the manner of answering each, complete with blackboard examples, in the best tradition of classroom instruction.

Concerning Question A, Ss were urged to rate the degree of student comprehension whenever possible, thus limiting the use of the can't tell category.

Instruction on answering Question B emphasized that a confidence rating should be given in all cases, even when Question A had been answered can't tell. In these cases, they were instructed to rate how confident they were that no cues for rating student comprehension were present.

Instruction on answering Question C consisted of contrasting "genuine effort" with attempting to give the impression of paying attention while actually thinking on unrelated topics. Our reason for asking Question C stemmed from the pretest sessions. Content analysis of the pretest session tapes revealed that judgments of student comprehension were based, in some cases, primarily on the raters' judgments of how hard the student in the clip was trying to understand. By pointing out the fact that these two factors should be considered independently, and requiring separate ratings of them, we hoped to reduce somewhat the variability in teacher ratings of student comprehension produced by confounding the two. These responses were not considered in the analysis of the results.

Following these instructions, and discussion of any questions which arose, two example film clips were shown. Following each, E marked duplicates of all

three rating scales at the blackboard. The example clips were shown with both visual image and sound track to all groups.

The *E* next described to each group the experimental treatment they were to receive. Then the first 10 clips of the experimental film were presented in appropriate form (visual image only, sound track only, or both). A 30-second interval was interposed after each clip during which *Ss* were to make their ratings. The room lights were turned off during the presentation of the clips, even in the S-O condition. Lighting was restored during each 30-second rating period.

At the end of the first 10 clips, room lights were left on and *E* suggested that *Ss* take a break and complete a questionnaire placed loose in the middle of the booklet. Information obtained on this questionnaire served to identify each *S* and check his placement in one of the experience-level groups. The rest of the questions served simply as filler material to keep *Ss* busy and prevent discussion of the experimental task. The *E* also explicitly requested that *Ss* not discuss their ratings during this period. Approximately 5 minutes were allowed each group for this break.

The second set of 10 clips was presented with no additional instructions. After the last clip, the lights were left on, booklets were collected, and *Ss* were dismissed.

RESULTS

Accuracy in Judging Student Comprehension

The main result of the experiment concerns the data supplied by the answers to Question A. Using the ratings preassigned by the investigators as an accuracy criterion, each *S* was given an accuracy score based on the number of clips he judged correctly. Table 4 presents the mean number of clips judged correctly in the nine cells of the experimental design. Analysis of variance reveals a highly significant experimental treatment effect ($F = 105.6, p < .001$). At all experience levels,

TABLE 4
MEAN NUMBER OF CLIPS RATED CORRECTLY:
QUESTION A

Group	Picture plus sound	Sound only	Picture only	Marginal
New interns	9.5	11.4	5.3	9.0
Old interns	10.8	11.4	5.2	9.0
Experienced teachers	10.3	10.2	4.8	8.6
Marginal	10.2	11.1	5.1	

TABLE 5
MEAN NUMBER OF CLIPS RATED ZERO:
QUESTION A, CAN'T TELL

Group	Picture plus sound	Sound only	Picture only	Marginal
New interns	3.0	4.6	1.0	3.0
Old interns	2.8	5.4	1.3	3.1
Experienced teachers	2.8	4.0	0.8	2.6
Marginal	2.9	4.7	1.1	

accuracy in the P + S and S-O conditions is significantly greater than would be expected on the basis of chance (for all *t*'s, $p < .05$ or less). There is no evidence to suggest that accuracy in the P-O condition is greater than chance. P + S and S-O means do not differ except among New Interns ($S-O > P + S, t = 2.60, p < .02$), and both reveal greater accuracy than that obtained in the P-O condition (for all *t*'s $p < .001$).

The analysis of variance and related *t* tests provide no support for our hypothesis that either standard teacher training or classroom experience in making judgments identical in content to those in the experiment improves the accuracy of these judgments. There are no differences in accuracy among experience groups which even approach statistical significance ($F = 1.2$). There is not even a significant interaction ($F = 1.2$) to suggest that experience exerts an influence in some of the experimental conditions.

Table 5 presents the mean number of clips rated can't tell in each cell of the design. Analysis of variance again reveals a very significant experimental treatment effect ($F = 45.1, p < .001$), and no experience level effect ($F = 1.4$) or interaction ($F < 1$). Inspection of Table 5 makes it quite clear that the condition in which *Ss* are least accurate (P-O) is the one in which the can't tell category is least often used. The authors feel that the can't tell category is frequently used in the S-O condition because there is a marked contrast between the clarity of verbal cues in most clips and the marked scarcity of verbal cues in the remaining clips. Consistent

with this hypothesis, the uniformity of cues in the clips in the P-O condition leads *S* to the feeling that if he rates amount of comprehension in one clip, he may as well do the same with all, or nearly all, clips. There is no apparent contrast in the amount of useful cues present in the clips when only the visual image is presented.

In reflecting on the strength of verbal cues as they affect accuracy of judgments, it became apparent to the investigators that a related factor—the latency between question and answer in the clip—is also a very important cue. Latency, as a basic cue for interpretation of degree of comprehension, includes not just the silent period preceding a student's reply but also the pauses and hesitations while the reply is being made. Thus there is both a verbal and a nonverbal component in latency of response.

To investigate the relationship between latency and judged student comprehension, the following analysis was performed. The latency between the first posing of a question by the instructor and the answering response by the student was determined for 14 of the 20 clips. The remaining six clips were not amenable to such an analysis since they did not contain a student response sufficiently specific to demarcate the question-answer sequence.

These 14 scorable clips were then arranged in order of their latencies, and dichotomized on that dimension. A Fisher exact test of the investigators' criterion ratings of the clips in these two latency categories results in $p = .0023$. In other words, the investigators' criterion judgments of student accuracy bears a strong inverse relationship to the student's latency in answering the question in the clip.

Confidence

An analysis of the confidence rating data was performed. Each *S* was given an average confidence score over all 20 clips. Table 6 presents the mean confidence in each of the nine cells of the design. Analysis of variance of these means reveals a significant experi-

TABLE 6
MEAN CONFIDENCE RATINGS: QUESTION B

Group	Picture plus sound	Sound only	Picture only	Marginal
New interns	3.59	3.73	2.94	3.38
Old interns	3.73	3.06	3.30	3.20
Experienced teachers	3.86	3.40	3.11	3.45
Marginal	3.54	3.38	3.09	

mental treatments effect ($F = 6.71$, $p < .01$), with no other significant effect. It is reassuring to see that where accuracy is not significantly greater than chance (the P-O condition), confidence in ratings is lower.

DISCUSSION

Clearly, visual feedback cues to live communicators in the one-to-many-to-one situation, e.g., the ordinary classroom, are not accurately interpreted. Yet verbal cues from students are usually either largely or totally absent in such situations. Furthermore, experience in the use of such visual cues, as in the case of the experienced classroom teacher, apparently does not improve the communicator's ability to interpret such cues accurately. These findings obtain in a situation where the communicator's only task is the observation and interpretation of these feedback cues. His attention is confined to one student, and he is not distracted by having to communicate simultaneously. It is obvious that in the live classroom setting where such distractions are many and the assessment of nonverbal cues must be done for many students more or less simultaneously, the task is a considerably more difficult one.

Since greatly increased verbal feedback is not feasible under normal classroom conditions, any method for substantially improving teacher accuracy in the interpretation of nonverbal feedback cues would be of value. The present investigators' next steps are therefore being directed towards that end.

(Received September 24, 1963)

STUDIES IN THE RELIABILITY AND VALIDITY OF THE CRITICAL INCIDENT TECHNIQUE

BENGT-ERIK ANDERSSON AND STIG-GÖRAN NILSSON

Institute of Education, University of Göteborg, Sweden

The critical incident technique was applied in analyzing the job of store managers in a Swedish grocery company. About 1800 incidents were collected, mainly by interviews, but also by questionnaires. Several reliability and validity aspects of the method were studied. When two-thirds of the incidents had been classified, 95% of the subcategories had appeared. The structure of the material was not influenced by the methods of collecting or by the interviewers. A repetition of the categorizing procedure was used to determine the stability of the subcategories. It was found that literature used in the training of personnel did not provide any additional, relevant information. Ratings of the subcategories support the assumption that the method covered the essential points in the job.

The Critical Incident Technique is a procedure used in the collection and analysis of incidents in which the holder of a position in a certain occupation has acted in a way which, according to some criterion, has been of decisive significance for his success or failure in a task. The method has been thoroughly described by Flanagan (1954) and by Andersson and Nilsson (1961, 1962). Although the method has been used in a practical manner in hundreds of job analyses, relatively little has been done to study the method itself with respect to either reliability or validity. This article describes some research which was undertaken in order to study such questions.

METHOD

Collection of Data

The incidents were collected at a Swedish grocery company which had a large number of branch

stores. The aim of the study was to determine the job and training requirements of store managers. Stores of the self-service type and the older traditional "groceries," "meat goods," and "mixed goods" shops were included in this study. The "traditional" shops did not have self-service. Critical incidents pertaining to the behavior of store managers were collected from four groups—superiors, store managers, assistants, and customers—by means of a questionnaire (only employee groups) and by individual interviews with all participants. The interviews took place during the spring of 1960. Table 1 gives the number of participants and the number of incidents reported by each of the different groups.

The employee groups were interviewed by five persons experienced in interview techniques, while the customers were interviewed by a group of students of psychology. The superiors were interviewed several times, both separately and in groups. These circumstances may partly explain the intergroup variations in the number of incidents reported.

The policy of the business enterprise was defined in very general terms. For that reason, the approval or disapproval of a certain behavior expressed by the person giving information must be taken as the only criterion of whether an incident was critical or not.

Classification of Incidents

In all 1,847 incidents were collected, 61% of which referred to units of successful or positive behavior. Store managers and assistants gave more positive than negative incidents (68% and 61%, respectively).

After classification of the incidents they could be grouped in 86 subcategories, 17 categories, and three major areas: A: Relation to customers, B: Relation to personnel, C: Relation to store and its sales. Area A deals with such behaviors as giving correct personal service, suggesting goods and giving information, and solving critical situations. Area B

TABLE 1

NUMBER OF PARTICIPANTS AND INCIDENTS COLLECTED

Group	N	Incidents
Superiors	10	289
Store managers		
Self-service	65	501
Traditional shops	57	312
Assistants		
Self-service	51	258
Traditional shops	49	201
Customers	178	286

contains behavior like creating job satisfaction, training the personnel, and cooperation. Incidents of knowledge of goods, advertising and displaying, planning and organizing are collected in Area C.

Table 3 shows that the superiors provided most incidents referring to store, which they probably found easiest to observe, the assistants most about personnel, and the customers most about service. The incidents from the store managers were more evenly distributed among all areas.

Table 4 gives rank correlations between the structure in the materials from the four groups of par-

TABLE 3

PERCENTAGE DISTRIBUTION OF INCIDENTS IN AREAS

Group	Area			Total
	A	B	C	
Superior	21.8	31.6	46.7	100.1
Store manager	29.9	34.7	35.3	99.9
Assistant	32.3	39.5	28.1	99.9
Customer	74.8	2.2	23.1	100.1

TABLE 2

PERCENTAGE DISTRIBUTION OF INCIDENTS IN CATEGORIES. CATEGORIES GROUPED IN AREAS

Area	Category	Percentage
A.	1. Gives customers correct personal service	7.4
	2. Suggests goods and gives information	5.4
	3. Satisfies customers' wishes	8.2
	4. Solves critical customer situations	15.2
B.	5. Creates job satisfaction	10.5
	6. Solves critical personnel problems	2.4
	7. Trains and advises personnel	12.3
	8. Takes part in personnel's work	4.2
	9. Cooperates with colleagues and superiors	1.0
C.	10. Advertising and display	7.2
	11. Is careful about appearances of the store	3.2
	12. Has good knowledge of goods	4.7
	13. Is careful with controls	1.9
	14. Plans and organizes	12.3
	15. Proposes rational solutions	1.5
	16. Shows practical skill	0.4
	17. Obeys and approves rules of the company	2.2
		100.0

Note.—Although all categories are positively formulated they also include the negative incidents.

ticipants calculated on the number of incidents in the categories. It appears from the table that there was a greater correlation between the employee groups than between them and the customers.

ANALYSIS

Saturation and Comprehensiveness

An important question in connection with the critical incident technique is whether or not the collection of data has been suf-

ficiently comprehensive to include all types of behavioral units that the method may be expected to cover. Only a few studies, however, have made systematic analyses of the collected material (see, e.g., Folley, 1953).

A first check of the material in the study reported here was made by classifying separately the last 215 of the incidents collected. It was found that all these incidents could be placed in the categories which had already been established. Following this a more detailed analysis was made by dividing all incidents obtained by means of interviews among the four groups of participants. All incidents from the same interviewee were placed together. Then the first 5% of incidents in each of the groups of participants were put together in one group and the next 5% of the incidents in another and so on. Twenty such groups were formed, numbered 1 to 20. In such a manner it was possible to determine how the number of subcategories increased with the number of collected incidents, i.e., how soon in the collection procedure the subcategories come up. This is shown in Table 5. The number of subcategories increased very rapidly at the beginning of the process of

TABLE 4

RANK CORRELATIONS BETWEEN THE MATERIALS FROM THE FOUR GROUPS OF PARTICIPANTS

	1	2	3	4
1. Superior		.68	.71	.44
2. Store manager			.87	.48
3. Assistant				.65
4. Customer				

Note.—Correlations of .48 and .61 are significant at the .05 and .01 levels, respectively.

TABLE 5
CUMULATIVE PERCENTAGE DISTRIBUTION OF
SUBCATEGORIES IN 5 PERCENTAGE GROUPS

	Group number						
	1	4	7	10	13	16	20
Percentage	48.8	81.4	88.4	91.9	94.2	95.3	100

collection. The increase soon became slower and when about two-thirds of the incidents had been classified 95% of all subcategories had appeared. Thus it is probable that the collection of data had not been stopped too early.

Reliability of Collecting Procedure

A question of concern in this study is whether, and to what extent, the number of incidents and their distribution in subcategories were affected by the methods of collection, and by the interviewers. The interviews provided five incidents per person (employee groups) and the questionnaire 2.5 incidents per person giving information. Tested by the Kolmogorov-Smirnov two-sample test the difference is significant. On the other hand, the structure of the two materials was not affected in the same way. The rank correlation between the sizes of the categories was .85, in spite of the fact that the percentage of replies to the questionnaire was very low (24%).

There were no great differences in the number of incidents per interview between the interviewers who interviewed the personnel. This was tested by the "Kruskal-Wallace one-way analysis of variance by ranks" for each of the following four employee groups: store managers in traditional shops and self-service and assistants in traditional shops and self-service. The only significant differences were between Interviewer A and Interviewers B, C, and E interviewing store managers in self-service. In these cases Interviewer A had nearly twice as many incidents per interview as the other interviewers.

Table 6 gives the rank correlations of the size of the categories between the materials of the interviewers. Interviewers D and E

conducted only 21 and 17 interviews while the other interviewers made, respectively 59, 81, and 44 interviews. The materials of Interviewers D and E are therefore very small, which may explain the lower correlation coefficients for these interviewers. Otherwise the structures of the materials obtained by the interviewers are very similar as shown by the coefficients of concordance (W) and the average correlations (r_{sav}), especially between Interviewers A, B, and C.

Control of Categorization

A much debated phase of the critical incident technique is categorization, which has been regarded as very subjective and difficult (Travers, 1958; Zaidenberg, 1953). It is clear that different people may systematize incidents in different ways. But one can always refer to the source material. The essential thing seems therefore to be that the category system chosen is an obvious one, and with as small a degree of arbitrariness and chance as possible.

The following experiment was made to ascertain whether other persons could easily place the incidents in the category system chosen. From each area two groups of 100 incidents each were taken at random. Each incident was written in two copies on special cards. Thus there were two different groups of incidents from each area, and each group could independently be classified twice.

TABLE 6
RANK CORRELATIONS BETWEEN THE INTERVIEWERS' MATERIALS (A-E), COEFFICIENTS OF CONCORDANCE (W) AND AVERAGE CORRELATIONS (r_{sav}) BETWEEN THE FIVE INTERVIEWERS' MATERIAL (A-E) AND BETWEEN INTERVIEWERS A, B, AND C (A-C)

	B	C	D	E
A	.84	.86	.76	.56
B		.81	.71	.46
C			.86	.75
D				.73
A-C		W = .89; r_{sav} = .83		
A-E		W = .78; r_{sav} = .72		

Note.—Rank correlations of .48 and .61 are significant at the .05 and .01 levels, respectively.

Twenty-four students of psychology took part in the experiment. Working in pairs they had to attempt to place a group of 100 incidents in corresponding categories. At their disposal they had the subcategory headings. It is possible to calculate the percentage agreement among those students which classified the same incidents (S-S) and also between the students and the original categorization (S-C). Table 7 gives the average proportion of incidents assigned by two groups to the same subcategories and categories for each area.

The first part of Table 7 shows that the agreement about some of the incidents is not especially high. The second part of the table, on the other hand, shows that even if two groups of students place an incident in different subcategories, there is a strong tendency to place the incident in the same category. It would appear that the agreement about the nature of the incident is stronger. This in turn suggests that the category system chosen is plausible and not too subjective.

Analysis of Contents of Training Literature

The question whether the critical incident method succeeded in including all the important aspects of work is connected primarily with validity. In order to study this problem, an analysis was made of the contents of the literature used by the enterprise in the internal training of managers (this is a continual training over years). The aim was to discover whether the contents could be fitted into the category system of the analysis. The study was only exploratory, and a rather rough measure was chosen as a unit, viz. one page of a book. There were more than 4,000 pages, 1,600 of which comprised an account of the history, aims, and organization of the enterprise and similar enterprises in the other Scandinavian countries. About 1,400 incidents could be placed in the subcategories and categories. Some 1,100 pages had no equivalents among the categories, but could be assigned to the areas.

In general it may be said that the analysis of contents did not reveal any new aspects. The literature that could not be assigned to the subcategories and categories was of such

TABLE 7

AVERAGE PERCENTAGE AGREEMENT AMONG STUDENTS (S-S) AND BETWEEN STUDENTS AND THE CRITERION (S-C), IN PLACING THE INCIDENTS IN SUBCATEGORIES AND CATEGORIES

Area	Proportion of incidents placed in the same			
	Subcategory		Category	
	S-S	S-C	S-S	S-C
A	68	61	80	75
B	64	62	81	81
C	66	64	82	85

Note.—Data given for each area separately.

a nature that it had no counterpart in the daily work of the store managers, but was intended to widen their knowledge and general education, and possibly prepare them for more qualified posts later.

The Importance of the Subcategories

Another aspect of validity that must be taken into consideration is whether the incidents collected are really critical in the meaning that a large number of judges find them important to the work in hand. Some critics consider the incidents so extreme and unique that they are of no practical importance. Only a few authors have tried to attack this problem systematically (Krumm, 1952; Vallance, Glickman, & Vasilas, 1954).

In order to study this, a rating form was constructed in which the 86 subcategories were to be rated on a six-point scale from 0 (unimportant) to 5 (of the greatest importance for the store manager's work). Forty-four superiors, 122 store managers, 45 assistants, and 89 students of psychology filled in the form. In order to neutralize eventual effects of fatigue about half of the raters started with the last subcategory. It was therefore possible to get a measure of the reliability of the rating form by correlating the median ratings for those raters who started at the beginning and those who started at the end. Since the result is based on both these ratings correction with Spearman-Brown's formula was done. Reliability coefficients were calculated for several rater

TABLE 8

RANK CORRELATIONS FOR THE SIX-POINT RATING
SCALE BETWEEN THE MEDIAN RATINGS
OF THE FOUR RATER GROUPS

	1	2	3	4
1. Superior		.88	.77	.58
2. Store manager			.85	.61
3. Assistant				.67
4. Student				

Note.—Correlations of .28 are significant at the .01 level.

groups. The average reliability coefficient was .83.

The number of subcategories with a median rating of below 3 (3 indicated that the subcategory in question was an important one) was, for the respective rater groups, 10, 8, 6, and 15. Only five subcategories had been rated as rather unimportant by all four groups. Thus it would seem that the method has revealed behavior units that may be considered important for the occupation with which we are concerned.

The rank correlations between the median ratings of the groups and the number of incidents that the corresponding groups of respondents have provided in the various subcategories showed a significant but slightly positive correlation. The correlations varied between .27 and .42. This means also that subcategories with few incidents may be important and that one must be careful not to regard frequency alone as a measure of the importance of a behavior unit.

If all the raters are pooled in one total group, Friedman's two-way analysis of variance reveals distinct differences between the relative importance of the subcategories within each category. Area C had most subcategories with high ratings, while Area A had the most with low ratings. Subcategories referring to certain items demanding knowledge, leadership, and encouragement of job satisfaction were rated highly. At the bottom were subcategories expressing opinions and attitudes, and some dealing with behavior units that were of a more unusual nature.

The students consistently gave lower rating figures than the other groups, probably because they were less involved in the enterprise

and the investigation, and were therefore less inclined to set very high value on the subcategories. Table 8 gives the rank correlations between the median ratings of the four rater groups. The same tendency appears there as in Table 4, that is, there is a greater correlation among the employee groups than between the employee groups and the customers or students.

CONCLUSIONS

The methodological checks of the critical incident technique that have been described in this article give a positive impression of the method.

The material collected seems to represent very well the behavior units that the method may be expected to provide. After a relatively small number of incidents had been classified, very few new behavior categories needed to be added.

The number and structure of the incidents were affected only slightly by different interviewers. Nor did the method of collecting the material affect the structure to any great extent, although fewer incidents were obtained with the questionnaire. Such findings may be compared with the studies of Finkle (1950) and Wagner (1948).

In addition the stability of the subcategory system appeared to be rather high when students tried to recategorize the incidents. It also seemed probable, according to content analysis of the training literature and the analysis of the questionnaire ratings, that the method covered the essential points in the job.

According to the results of the studies reported here on the reliability and validity aspects of the critical incident technique, it would appear justifiable to conclude that information collected by this method is both reliable and valid.

REFERENCES

ANDERSSON, B.-E., & NILSSON, S.-G. Critical incident metoden för analys av arbets- och utbildningskrav. Reports, University of Göteborg, Institute of Education, 1961. (Mimeo)

ANDERSSON, B.-E., & NILSSON, S.-G. An application of the critical incident technique to the study of job and training requirements of shop managers.

- Reports, University of Göteborg, Institute of Education, 1962. (Mimeo)
- FINKLE, R. B. A study of the critical requirements of foremanship. *Univer. Pittsburgh Bull.*, 1950, **46**, 291-297. (Abstract)
- FLANAGAN, J. C. The critical incident technique. *Psychol. Bull.*, 1954, **51**, 327-358.
- FOLLEY, J. D., JR. Development of a list of critical requirements for retail sales personnel from the standpoint of customer satisfaction. Unpublished master's thesis, University of Pittsburgh, 1953.
- KRUMM, R. L. Critical requirements of pilot instructors. *USAF Hum. Resour. Res. Cent. tech. Rep.*, 1952, No. 52-1.
- TRAVERS, M. W. *An introduction to educational research*. New York: Macmillan, 1958.
- VALLANCE, T. R., GLICKMAN, A. S., & VASILAS, J. N. Critical incidents in junior officer duties aboard destroyer-type vessels. *USN Bur. Naval Personnel tech. Bull.*, 1954, No. 54-4.
- WAGNER, R. F. A group situation compared with individual interviews for securing personnel information. *Personnel Psychol.*, 1948, **1**, 93-107.
- ZAIDENBERG, D. L'étude du travail chez Flanagan. *Bull. Cent. Etud. Rech. Psychotech.*, 1953, **2**, 2-12.

(Received September 26, 1963)

14

